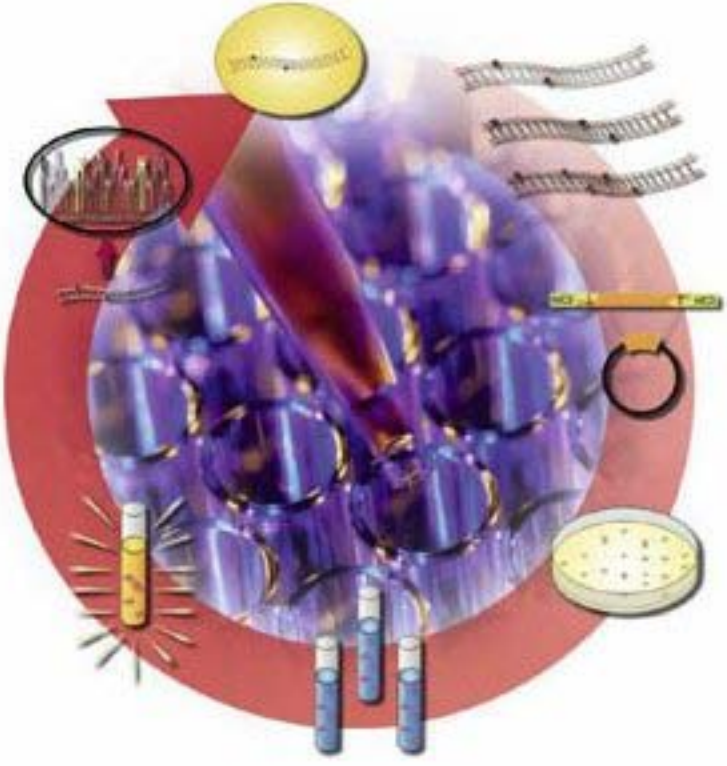
Edited by Susanne Brakmann,
Andreas Schwienhorst

WILEY-VCH

Evolutionary Methods in Biotechnology

Clever Tricks for Directed Evolution



VISIT…

Evolutionary Methods in Biotechnology

Edited by Susanne Brakmann and Andreas Schwienhorst

Further Titles of Interest

G. Kahl
The Dictionary of Gene Technology
3rd edition 2004, ISBN 3-527-30765-6

A.S. Bommarius, B.R. Riebel
Biocatalysis – Fundamentals and Applications
2004, ISBN 3-527-30344-8

T. Dingermann, D. Steinhilber, G. Folkers (eds.)
Molecular Biology in Medicinal Chemistry
2003, ISBN 3-527-30431-2

S. Brakmann, K. Johnsson (eds.)
Directed Molecular Evolution of Proteins
2002, ISBN 3-527-30423-1

V. Braun, F. Götz (eds.)
Microbial Fundamentals of Biotechnology
2001, ISBN 3-527-30615-3

Evolutionary Methods in Biotechnology

Clever Tricks for Directed Evolution

Edited by Susanne Brakmann
and Andreas Schwienhorst

WILEY-VCH Verlag GmbH & Co. KGaA

Dr. Susanne Brakmann
Institut für Zoologie
Angewandte Molekulare Evolution
Universität Leipzig
Liebigstr. 18
04103 Leipzig, Germany
sbrakma@uni-leipzig.de

Dr. Andreas Schwienhorst
Institut für Mikrobiologie und Genetik
Universität Göttingen
Grisebachstr. 8
37077 Göttingen, Germany
aschwie1@gwdg.de

This book was carefully produced. Nevertheless, authors, editors and publisher do not warrant the information contained therein to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No. applied for.

British Library Cataloguing-in-Publication Data: A catalogue record for this book is available from the British Library

Die Deutsche Bibliothek – CIP Cataloguing-in-Publication-Data: A catalogue record for this publication is available from Die Deutsche Bibliothek

ISBN 3-527-30799-0

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Printed on acid-free paper
All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into a machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

Composition: EDV-Beratung Frank Herweg, Leutershausen
Printing: betz-druck gmbh, Darmstadt
Bookbinding: J. Schäffer GmbH & Co. KG, Grünstadt

Printed in the Federal Republic of Germany.

Contents

1 Introduction
Susanne Brakmann and Andreas Schwienhorst 1
References 3

2 Generation of Mutant Libraries Using Random Mutagenesis
Susanne Brakmann and Björn F. Lindemann 5
2.1 Introduction 5
2.2 Materials 6
2.2.1 Materials for Random PCR Mutagenesis 6
2.2.2 Materials for Mutator Strain Passage 6
2.3 Protocols 7
2.3.1 Protocol for Random PCR Mutagenesis According to Joyce 7
2.3.2 Protocol for Mutator Strain Passage 8
2.4 Troubleshooting 10
References 11

3 DNA Shuffling
Hikaru Suenaga, Masatoshi Goto, and Kensuke Furukawa 13
3.1 Introduction 13
3.2 Materials 15
3.2.1 For Preparation of Parental Genes 15
3.2.2 For Random Fragmentation by DNase I 15
3.2.3 For Collection of DNA Fragments in Specific Molecular Size Ranges 16
3.2.4 For Reassembly of These Fragments by Primerless PCR . 16
3.2.5 For Amplification of Reassembled Products by Conventional PCR with Primers 16
3.3 Protocol 17
3.3.1 Preparation of Parental Genes 17
3.3.2 Random Fragmentation by DNase I 17
3.3.3 Collection of DNA Fragments in Specific Molecular Size Ranges 18
3.3.4 Reassembly of These Fragments by Primerless PCR . . . 19
3.3.5 Amplification of Reassembled Products by Conventional PCR with Primers 20

3.4 Troubleshooting 21
3.4.1 Insufficient DNase I Fragmentation 21
3.4.2 Little or No Product of Primerless PCR 21
3.4.3 Little or No Product of PCR with Primers 21
3.4.4 The Product of PCR with Primers is Multi-banded 22
3.5 Amplification Examples 22
References 23

4 DNA Recombination Using StEP
Milena Ninkovic 25
4.1 Introduction 25
4.2 Materials 26
4.2.1 StEP PCR 26
4.2.2 Purification of an Appropriate DNA Fragment 27
4.2.3 Equipment 27
4.3 Protocol 27
4.4 Technical Tips 28
4.4.1 Problem: Little or No PCR Product (Full-length Product) after PCR 28
4.4.2 Problem: High Background Levels of DNA after PCR . . 28
4.5 StEP in Directed Evolution 29
References 30

5 FACS Screening of Combinatorial Peptide and Protein Libraries Displayed on the Surface of *Escherichia coli* Cells
Thorsten M. Adams, Hans-Ulrich Schmoldt, and Harald Kolmar 31
5.1 Introduction 31
5.2 Materials 35
5.2.1 *Escherichia coli* Strains and Plasmids 35
5.2.2 Liquid Media and Agar Plates 35
5.2.3 Biological and Chemical Materials 36
5.2.4 Equipment 36
5.3 Protocols 36
5.3.1 Verification of Cell Surface Exposure of the Passenger Protein 36
5.3.2 Labeling of the Target Protein 37
5.3.3 Library Construction 37
5.3.4 Combinatorial Library Screening by FACS and MACS . . 40
5.4 Troubleshooting 42
5.5 Major Applications 44
References 44

6 Selection of Phage-displayed Enzymes
Patrice Soumillion 47
6.1 Introduction 47
6.2 Materials 50

6.2.1 Buffers, Reagents and Consumables 50
6.2.2 Strains and Vectors 50
6.3 Protocols 51
6.3.1 The Phage-enzyme 51
6.3.2 Library Construction 55
6.3.3 Selection 59
6.4 Troubleshooting 62
6.4.1 Phage Titers Are Not Reproducible 62
6.4.2 Phage-enzymes Degrade with Time 63
6.4.3 Phages Are Not Genetically Stable 63
6.4.4 The 'out/in' Ratio Does Not Increase with Selection Rounds 63
6.5 Major Applications 63
References 64

7 Selection of Aptamers
Heiko Fickert, Heike Betat, and Ulrich Hahn 65
7.1 Introduction 65
7.2 Materials 66
7.2.1 Immobilization of Target Molecules 66
7.2.2 PCR 67
7.2.3 *In vitro* Transcription 67
7.2.4 RNA Purification 67
7.2.5 Selection of Aptamers 67
7.2.6 Reverse Transcription 68
7.3 Protocols 68
7.3.1 Selection of RNA Aptamers 68
7.3.2 Selection of 2′-Modified RNA Aptamers 75
7.3.3 Selection of ssDNA Aptamers 76
7.3.4 Cloning and Sequencing 77
7.3.5 Characterization of Aptamers 77
7.3.6 Example: Isolation of Moenomycin A-specific Aptamers 79
7.4 Troubleshooting 82
7.5 Major Applications 83
References 83

8 Methods for Selecting Catalytic Nucleic Acids
Benjamin L. Holley, and Bruce E. Eaton 87
8.1 Introduction 87
8.2 Materials and Equipment 88
8.3 Protocols 91
8.3.1 Generating the Starting Library 91
8.3.2 Transcription 97
8.3.3 Ligation 99
8.3.4 Nucleic Acid-catalyzed Reactions 102
8.3.5 Reverse Transcription 104

8.3.6 Partitioning 105
8.4 Troubleshooting 108
8.5 Major Applications 109
References 109

9 High-throughput Screening of Enantioselective Industrial Biocatalysts
Manfred T. Reetz 113
9.1 Introduction 113
9.2 Materials and Equipment 115
9.2.1 Assays Based on Mass Spectrometry 115
9.2.2 Assays Based on NMR Spectrometry 116
9.2.3 Assay Based on FTIR Spectroscopy 116
9.2.4 Assays Based on UV/Visible Spectroscopy 116
9.2.5 Enzyme-coupled UV/Visible-based Assay for Hydrolases 117
9.3 Protocols 117
9.3.1 Assays Based on Mass Spectrometry 117
9.3.2 Assays Based on NMR Spectroscopy 121
9.3.3 Assay Based on FTIR Spectroscopy 125
9.3.4 Assays Based on UV/Visible Spectroscopy 129
9.3.5 Enzyme-coupled UV/Visible-based Assay for Hydrolases 132
9.3.6 Further Assays 133
9.4 Troubleshooting 138
9.4.1 Comments on the Kazlauskas Test 138
9.4.2 Potential Problems when Performing Kinetic Resolution . 138
9.5 Conclusions 139
References 139

10 Computer-assisted Design of Doped Libraries
Dirk Tomandl and Andreas Schwienhorst 143
10.1 Introduction 143
10.2 Materials 146
10.3 Protocol 147
10.4 Troubleshooting 150
10.5 Major Applications 150
References 151

11 Directed *in silico* Mutagenesis
Markus Wiederstein, Peter Lackner, Ferry Kienberger, Manfred J. Sippl . . . 153
11.1 Introduction 153
11.2 Materials 155
11.2.1 PDB Files 155
11.2.2 Knowledge-based Potentials 156
11.2.3 Polyprotein, Z-scores 160
11.2.4 *In silico* Mutagenesis 162
11.2.5 Summary 163

11.3 Protocol . . . 163
11.3.1 ProSa Setup and Interaction . . . 163
11.3.2 ProSa Objects . . . 164
11.3.3 Session 1 (`mut_script1.cmd`) . . . 164
11.3.4 Session 2 (`mut_script2.cmd`) . . . 166
11.3.5 Session 3 (`mut_script3.cmd`) . . . 168
11.3.6 Session 4 (`mut_script4.cmd`) . . . 170
11.3.7 Tips & Tricks . . . 171
11.4 Troubleshooting . . . 173
11.5 Major Applications . . . 174
References . . . 175

12 RNA Folding *in silico*
Christoph Flamm, Ivo L. Hofacker, Peter F. Stadler . . . 177
12.1 Introduction . . . 177
12.2 Materials . . . 178
12.2.1 Typographical Conventions . . . 179
12.2.2 RNA Web Services . . . 180
12.3 Protocols . . . 181
12.3.1 Secondary Structures for Individual Sequences . . . 181
12.3.2 Consensus Structures of a Sample of Sequences . . . 183
12.3.3 Sequence Design . . . 184
12.3.4 Analysis of SELEX Experiments . . . 186
12.3.5 A Note for the Experts: Write your Own RNA Programs . . . 187
12.4 Troubleshooting . . . 187
12.5 Caveats . . . 188
References . . . 189

13 Patenting in Evolutionary Biotechnology
Martina Leimkühler and Hans-Wilhelm Meyers . . . 191
13.1 Introduction . . . 191
13.2 The Nature of Patents . . . 191
13.3 What Can Be Patented . . . 192
13.4 The Requirement of Novelty . . . 193
13.5 The Requirement of Inventiveness . . . 196
13.6 The Requirement of Utility . . . 197
13.7 The Requirements of Enablement and Written Description . . . 197
13.8 Patent Prosecution . . . 199
13.9 Search Tools . . . 204
13.10 The First-to-invent Principle of the United States and Its Consequences on Laboratory Notebook Keeping . . . 206
13.11 Summary . . . 209
References . . . 209

Subject Index . . . 211

List of Contributors

Thorsten M. Adams
Institut für Mikrobiologie und Genetik
Universität Göttingen
Grisebachstr. 8
37077 Göttingen
Germany

Heike Betat
Max-Planck-Institut für
Evolutionäre Anthropologie
Deutscher Platz 6
04103 Leipzig
Germany

Susanne Brakmann
Institut für Zoologie
Angewandte Molekulare Evolution
Universität Leipzig
Liebigstr. 18
04103 Leipzig
Germany

Bruce E. Eaton
Department of Chemistry
North Carolina State University
Raleigh, NC 27695
USA

Heiko Fickert
Universität Hamburg
Institut für Biochemie
und Lebensmittelchemie
Martin-Luther-King Platz 6
20146 Hamburg
Germany

Christoph Flamm
Institut für Theoretische Chemie
und Strukturbiologie
Universität Wien
Währingerstr. 17
1090 Wien
Austria

Kensuke Furukawa
Department of Bioscience
and Biotechnology
Kyushu University
Fukuoka 812-8581
Japan

Masatoshi Goto
Department of Bioscience
and Biotechnology
Kyushu University
Fukuoka 812-8581
Japan

Ulrich Hahn
Universität Hamburg
Institut für Biochemie
und Lebensmittelchemie
Martin-Luther-King Platz 6
20146 Hamburg
Germany

Ivo L. Hofacker
Institut für Theoretische Chemie
und Strukturbiologie
Universität Wien
Währingerstr. 17
1090 Wien
Austria

Benjamin L. Holley
Department of Chemistry
North Carolina State University
Raleigh, NC 27695
USA

Ferry Kienberger
Center of Applied Molecular
Engineering
Institute of Chemistry
and Biochemistry
University of Salzburg
Jakob Haringerstr. 5
5020 Salzburg
Austria

Harald Kolmar
Institut für Mikrobiologie und Genetik
Universität Göttingen
Grisebachstr. 8
37077 Göttingen
Germany

Peter Lackner
Center of Applied Molecular
Engineering
Institute of Chemistry
and Biochemistry
University of Salzburg
Jakob Haringerstr. 5
5020 Salzburg
Austria

Martina Leimkühler
Evotec OAI AG
Schnackenburgallee 114
22525 Hamburg
Germany

Björn Lindemann
Bioagency AG
Schnackenburgallee 116a
22525 Hamburg
Germany

Hans-Wilhelm Meyers
Deichmannhaus am Dom
Bahnhofsvorplatz 1
50667 Köln
Germany

Milena Ninkovic
Max-Planck-Institut
für Experimentelle Medizin
Hermann-Rein-Str. 3
37075 Göttingen
Germany

Manfred T. Reetz
Max-Planck-Institut
für Kohlenforschung
Kaiser-Wilhelm-Platz 1
45470 Mülheim/Ruhr
Germany

Hans-Ulrich Schmoldt
Institut für Mikrobiologie und Genetik
Universität Göttingen
Grisebachstr. 8
37077 Göttingen
Germany

Andreas Schwienhorst
Institut für Mikrobiologie und Genetik
Universität Göttingen
Grisebachstr. 8
37077 Göttingen
Germany

Manfred J. Sippl
Center of Applied Molecular
Engineering
Institute of Chemistry
and Biochemistry
University of Salzburg
Jakob Haringerstr. 5
5020 Salzburg
Austria

Patrice Soumillion
Institut des Sciences de la Vie
Université Catholique de Louvain
Place Pasteur 1/1b
1348 Louvain-la-Neuve
Belgium

Peter F. Stadler
Institut für Informatik
Universität Leipzig
Kreuzstrasse 7b
04103 Leipzig
Germany

Hikaru Suenaga
Institute for Biological Resources
and Functions
National Institute of Advanced
Industrial Science and Technology
Tsukuba, Ibaraki 305-8566
Japan

Dirk Tomandl
Graffinity Pharmaceuticals AG
Im Neuenheimer Feld 518 - 519
69120 Heidelberg
Germany

Markus Wiederstein
Center of Applied Molecular
Engineering
Institute of Chemistry
and Biochemistry
University of Salzburg
Jakob Haringerstr. 5
5020 Salzburg
Austria

1 Introduction

Susanne Brakmann and Andreas Schwienhorst

Since the landmark papers of Manfred Eigen [1, 2] and Sol Spiegelman [3, 4], the concept of Darwinian evolution has had a major impact on the design of biomolecules with tailored properties. 'Directed evolution', 'applied evolution', and 'evolutionary biotechnology' are different expressions that all describe an 'evolutionary' type of optimization strategy that comprises several cycles each consisting of (1) molecular library preparation to create the desired molecular diversity, (2) functional selection or screening, and (3) error-prone amplification or chemical modification of selected species to generate a new library of molecules (Fig.1.1). The ultimate goal is to identify molecular species that are well-adapted to a given profile of defined demands. Biocatalysts, for example, may be generated to exhibit high processivity, enantioselectivity, or tolerance to high temperatures or organic solvents.

The book presented here is intended as a practical state-of-the-art compilation of methods related to the topic of directed evolution and hence is complementary to the recent successful book *Directed Molecular Evolution of Proteins* [5]. The methods are described in sufficient detail to serve as 'recipes' in a 'cookbook'. They are easy to follow by laboratory staff, from the technical assistant to the postdoctoral academic or industrial specialist.

The sequence of chapters mirrors the steps in a standard directed-evolution experiment. In the beginning, various methods for the creation of molecular diversity are considered. S. Brakmann and B.F. Lindemann (Chapter 2) present protocols for the generation of mutant libraries by random mutagenesis. Two chapters deal with the particularly powerful approach of *in-vitro* recombination. H. Suenaga, M. Goto, and K. Furukawa (Chapter 3) describe the application of DNA shuffling, and M. Ninkovic (Chapter 4) presents DNA recombination by the StEP method.

Next, several chapters are concerned with techniques of selection and/or mass screening technologies. T. Adams, H.-U. Schmoldt, and H. Kolmar (Chapter 5) describe the FACS-based screening of combinatorial peptide and protein libraries. P. Soumillion (Chapter 6) presents some of the latest developments in the selection of phage-displayed enzymes. In Chapter 7, H. Fickert, H. Betat, and U. Hahn provide methods for the selection of specific target-binding nucleic acids, i. e., aptamers. Related methods for the generation of catalytic nucleic acids are described by B.L. Holley and B.E. Eaton (Chapter 8). The part on functional selection and screening closes with a description of high-throughput screening approaches, in

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst
Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-30799-0

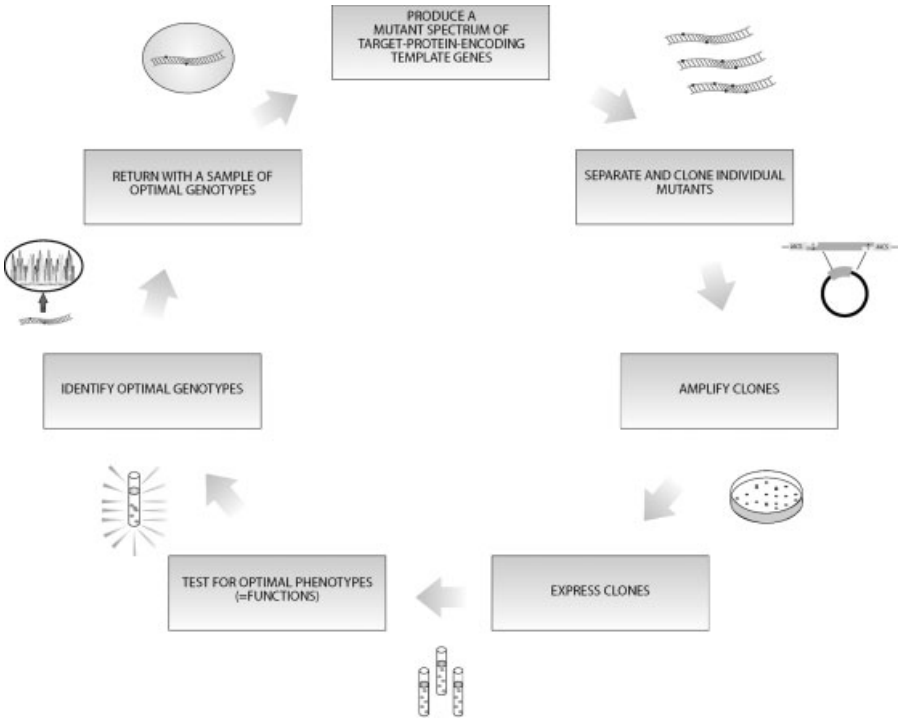

PRODUCE A MUTANT SPECTRUM OF TARGET-PROTEIN-ENCODING TEMPLATE GENES
SEPARATE AND CLONE INDIVIDUAL MUTANTS
AMPLIFY CLONES
EXPRESS CLONES
TEST FOR OPTIMAL PHENOTYPES (=FUNCTIONS)
IDENTIFY OPTIMAL GENOTYPES
RETURN WITH A SAMPLE OF OPTIMAL GENOTYPES

Fig. 1.1. Scheme of directed evolution. Starting from a pool of mutant genes, single clones are expressed and their phenotype is evaluated in a selection or screening step. Clones with desired phenotypes provide genes that are the basis for the subsequent cycle of selection.

particular, to produce enantioselective industrial biocatalysts, provided by M.T. Reetz (Chapter 9).

Combinatorial mutagenesis easily produces a degree of molecular diversity that far exceeds the number of different proteins or functional nucleic acids that can be produced in a single experiment. As the number n of randomized amino acid positions in a protein grows, the number of possible combinations increases as 20^n. Hence, complete coverage of a library with 9 randomized positions requires a library size well above 10^{11} molecules. Since in a standard random library, functional molecules are usually highly diluted in a large background of nonfunctional, e. g., misfolded, molecules, it may be meaningful to restrict variations to a certain subset of promising molecules. Three chapters deal with theoretical computer-based methods to predict these promising molecular species. D. Tomandl and A. Schwienhorst (Chapter 10) report a 'doping' algorithm that helps to design random codons for only subsets of amino acids, at the same time minimizing stop codons. M. Wiederstein, P. Lackner, F. Kienberger, and M.J. Sippl (Chapter 11) provide algorithms to predict (mutant) protein structures as a means of *in silico* mutagenesis, e. g., to enhance the probability of generating properly folded mutant proteins. C. Flamm, I.L.

Hofacker, and P.F. Stadler (Chapter 12) pursue a similar goal concerning functional nucleic acids and provide various *in silico* tools to predict RNA folding.

In the past 10 years, directed evolution has gained considerable attention as a commercially important strategy for rapidly designing molecules with properties tailored for the biotechnological and pharmaceutical market. Therefore, legal protection of methods and molecules has become an important issue. Hence, the book closes with Chapter 13, by M. Leimkühler and H.W. Meyers on patenting issues in evolutionary biotechnology.

Since the first evolution experiments by Sol Spiegelman, Manfred Eigen, and coworkers, the field of directed evolution itself has evolved into a plethora of different methodologies that can hardly be covered comprehensively in a standard textbook. We nevertheless tried to provide a collection of protocols useful to the novice as well as to the scientist experienced in the field. We hope to provide a practical starting point and at the same time inspire scientists to develop their own variations on the evolutionary theme.

We thank all the authors for their contributions, and Peter Gölitz and Frank Weinreich of Wiley-VCH for their help in publishing this book.

References

1. Eigen, M. *Die Naturwissenschaften*, **1971**, 58, 465–523.
2. Eigen, M. and Gardiner, W. *Pure Appl. Chem.*, **1984**, 56, 967–976.
3. Spiegelman, S., Haruna, I., Holland, I.B., Beaudreau, G., and Mills, D. *Proc. Natl. Acad. Sci. USA*, 1965, 54, 919-927.
4. Mills, D.R., Peterson, R.L., and Spiegelman, S. *Proc. Natl. Acad. Sci. USA*, **1967**, 58, 217–224.
5. Brakmann, S., Johnsson, K., eds. *Directed Molecular Evolution of Proteins: or How to Improve Enzymes for Biocatalysis*, Wiley-VCH, Weinheim, **2002**.

2 Generation of Mutant Libraries Using Random Mutagenesis

Susanne Brakmann and Björn F. Lindemann

2.1 Introduction

Engineering enzymes by applying directed evolution strategies involves the generation of molecular libraries that are as large and as diverse as possible. However, mutant libraries of enzymes, which usually consist of more than 100 amino acids, are inaccessible by automatic chemical synthesis. These are better available by mutagenesis at the nucleotide level. During the past decade, a series of experimental strategies has been developed for generating DNA mutant libraries that differ in diversity, that is, in the extent of sequence space covered, and in their way to deal with complex libraries.

Random mutagenesis is a widespread strategy which targets whole genes. This may be achieved by passing cloned genes through mutator strains [1, 2], by treating DNA or whole bacteria with various chemical mutagens [3–6], or by "error–prone" [7, 8] or "hypermutagenic" PCR [9]. Due to its simplicity and versatility, random PCR mutagenesis has emerged as the most common technique and can result in mutation frequencies as high as 10% per nucleotide position. The incorporation of nucleotide analogs that promote base pair mismatching during PCR has even been found to cause overall mutation frequencies of up to 19% per position and PCR [10]. With alterations of some PCR conditions, the mutation rate may be adjusted to the appropriate level (see Table 2.2). Usually, a maximal number of mutants (and no wildtype) is required, of which as many variants as possible should be active. For example, catalytically active variants of enzymes like HIV reverse transcriptase, *Taq* polymerase, or HSV–1 thymidine kinase almost never contain more than five amino acid substitutions [11]. We should also mention that the number of amino acid substitutions accessible by error–prone PCR is limited, because on the one hand, the reaction may bias the distribution of mutation type (depending on the sequence), and on the other hand, multiple substitutions within a single codon are extremely rare.

Alternative random mutagenesis strategies have been developed for targeting single or a few amino acids or selected regions of a protein that might be important for a certain function. By focusing on only the positions of interest and their close environment or by reducing the set of amino acids per randomized position (see Chapter 10 by Tomandl), the library size can be drastically reduced. Typically,

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst
Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-30799-0

randomization of small gene fragments is achieved by substituting a wildtype gene fragment with a synthetic oligonucleotide which contains random positions or regions (random cassettes [12, 13]) or semi–random ranges (spiked oligonucleotides [14]). Randomization of defined positions or regions is achieved with automatic solid-phase DNA synthesis, by programming the desired International Union of Biochemistry (IUB) mix codes. The introduction of stop codons can be reduced by allowing only G and C (IUB mix code: S) at the third position of each codon. Complete permutation of a single amino acid position may thus enable finding nonconservative replacements that are inaccessible by random point mutagenesis.

In this chapter, two approaches are described for the introduction of random point mutations into whole genes: (1) PCR mutagenesis and (2) mutator strain passage. Both procedures involve the cloning of target genes into custom plasmid vectors ready for the functional expression of enzyme variants. Alternatively, mutant gene libraries may be expressed by using commercially available *in vitro* transcription/translation systems. However, this topic is not discussed here.

2.2 Materials

2.2.1 Materials for Random PCR Mutagenesis

1. Template DNA encoding the gene of interest.
2. Oligonucleotide primers containing the desired restriction sites for cloning.
3. Expression vector with suitable promotor, multiple cloning site, and fusion tag, where applicable (e.g., six-histidine tag).
4. *Taq* DNA polymerase and buffer.
5. Deoxynucleoside triphosphates (10 mM each).
6. $MnCl_2$ (100 mM).
7. $MgCl_2$ (100 mM).
8. PCR and gel purification (spin) kit.
9. Agarose gel electrophoresis equipment.
10. Restriction endonucleases, alkaline phosphatase, T4 DNA ligase.
11. Competent *E. coli* cells (high quality is required; $\geq 10^9$ transformants/μg supercoiled DNA).
12. Luria Bertani (LB) media and appropriate antibiotic.

2.2.2 Materials for Mutator Strain Passage

1. Plasmid vector encoding the target gene in a genetic context ready for expression in *E. coli*.
2. Mutator strain: XL1–Red (*mutD, mutS, mutT*; Stratagene).

3. Plasmid preparation (spin) kit.
4. Amplification strain: XL1–Blue (Stratagene).
5. LB media and appropriate antibiotic.

2.3 Protocols

2.3.1 Protocol for Random PCR Mutagenesis According to Joyce

A series of parameters is used to substantially increase the overall error frequency of *Taq* DNA polymerase. This enzyme lacks 3′-5′ exonuclease activity and exhibits an error rate of 0.8–1.1 $\times 10^{-4}$ base substitutions/bp of product under standard conditions [15, 16]. The mutagenic PCR conditions include (1) increased Mg^{2+} concentration for stabilizing noncomplementary base pairs [17], (2) the addition of Mn^{2+} for reducing the base pairing specificity [18], (3) unbalanced dNTP stoichiometry for forcing misincorporation [7], and (4) increased polymerase concentration for enhancing the probability of elongation of misprimed termini [19]. The protocol below largely follows the procedure originally conceived by G. Joyce [8, 20]:

1. Prepare a 10X dNTP mix consisting of 2 mM each of dATP and dGTP and 10 mM each of dCTP and dTTP.
2. Setup a PCR reaction starting with 0.05–0.2 pmol of template DNA, 50 pmol of each primer, 10 µL 10X PCR buffer, 10 µL 10X dNTP mix, 0.5 mM $MnCl_2$, 5 U *Taq* DNA polymerase, and water to a final volume of 100 µL. The manganese solution should be added just prior to the polymerase (see section 2.4, note 1).
3. Perform PCR cycling following the standard conditions for this template/primer system.
4. Analyze 5 µL of the reaction on an 0.8% agarose gel (see section 2.4, notes 2–4). Usually, the yield of an error-prone PCR reaction is lower than that of a standard PCR; however, one 100 µL reaction will yield $\approx$ 1–2 µg of crude PCR product (10^{10}–10^{11} molecules). For efficient cloning, 2–5 100-µL reactions should be prepared.
5. Purify the product using a PCR purification (spin) kit.
6. Digest PCR product and vector according to standard protocols [21]. Dephosphorylate the vector using alkaline phosphatase and purify the DNA by agarose gel electrophoresis.
7. Ligate vector and insert, applying at least 3-fold molar excess of the insert (PCR product).
8. Transform competent *E. coli* according to the supplier's manual and cultivate in LB media (plates or liquid cultures, depending on the selection/screening approach; see section 2.4, note 5).

Using DNA fragments of various origins, nucleotide compositions, and lengths (maximum of $\approx$3 kb), we observed mutation frequencies of 0.93 $\pm$ 0.06% per

Table 2.1. Sequence context of mutation types and their frequencies observed after application of error-prone PCR as described herein and sequencing of cloned genes. Targets: T7 RNAP, coding sequence of T7 RNA polymerase, PolI, coding sequence of *E. coli* DNA polymerase I, Intron, cDNA of *Tetrahymena thermophila* intron.

Mutation type	T7 RNAP	PolI	Intron
Transversion (G ⇌ C, A ⇌ T)	32.2%	45.7%	37.0%
Transition (G:C ⇌ A:T)	58.0%	45.4%	61.1%
Deletions	9.8%	6.2%	1.2%
Insertions	–	2.8%	–

position over the total course of an error-prone PCR. The most abundant changes were single-base substitutions (see Table 2.1). Depending on the sequence, we observed that the reaction biases the distribution of mutation types in favor of transitions (A:T ⇌ G:C changes) in two of three instances.

We observed only minor influence of the sequence context – except with the *Tetrahymena thermophila* intron fragment. The intron mutants exhibited a series of base substitutions (12%) that occurred pairwise, thus yielding new complementary base pairs within the strong secondary structural elements of the ribozyme. This observation was surprising and might be due to the fact that the intron was cloned as part of the *lacZ* gene, such that the correctly spliced pre-mRNA restored the *lacZ* reading frame. Mutant clones exhibiting a higher splicing efficiency might have been selected because of their elevated galactosidase activity.

Although 10% of all PCR–generated variants contained frameshift mutations (leading to early termination of translation), the mutation frequency of 0.93% on the nucleotide level yielded 1.78% amino acid substitutions after translation. Because multiple substitutions within a single codon are extremely rare, a proportion of 30% of all mutations were silent; that is, they did not influence the amino acid sequence.

For some purposes, an adjustment of the mutation rate to lower or higher values might be desirable. Table 2.2 summarizes the effects that were observed after alteration of selected PCR conditions.

In principle, the product of a mutagenic PCR may be used for successive rounds of random mutagenization. Therefore, a small aliquot of the first reaction may be employed to seed the next one. We found that it is necessary to gel-purify the product of the first PCR before proceeding with another amplification cycle. Otherwise, the PCR yielded exclusively nonspecific amplification products.

2.3.2 Protocol for Mutator Strain Passage

As an alternative to random mutagenesis by error-prone PCR, expression vectors containing the gene of interest may be propagated in mutator *E. coli* strains like XL1–Red (Stratagene). This strains contains mutations in three DNA repair pathways and exhibits a more than 5000–fold increased spontaneous mutation rate (3.5×10^{-6}), compared to wildtype *E. coli* (7×10^{-10}) [22]. Provided that all mu-

Table 2.2. Effects of PCR conditions on mutation types and frequencies using a 270-bp fragment of the lacZ gene as a template. Mean values per sequenced clone are given. The mutation rate (in percent per position and PCR) includes all mutation types.

Protocol	**A**	**B**	**C**	**D**	**E**	**F**
$MgCl_2$ [mM]	8.5	6.5	6.5	6.5	6.5	6.5
$MnCl_2$ [mM]	0.2	0.5	0.5	0.5	0.5	0.5
dATP [μM]	40	20	4	200	20	20
dGTP [μM]	40	20	200	20	200	20
dCTP [μM]	200	200	200	200	200	200
dTTP [μM]	200	200	200	200	200	200
dITP [μM]	-	200	-	-	-	-
No. of sequenced clones	4	4	10	6	6	5
Deletions	0.75	1.25	0.2	0.67	0.5	–
Insertions	–	–	–	–	–	–
C→T	0.75	0.5	0.1	2.0	0.33	0.2
T→C	0.75	0.5	10.0	0.17	1.17	2.2
A→G	0.25	1.5	6.7	0.17	1.83	1.0
G→A	0.75	0.25	-	1.33	0.17	0.4
C→A	0.25	-	0.2	0.33	0.17	0.2
T→A	1.25	1.75	0.4	0.83	1.17	1.0
C→G	–	0.25	0.1	0.17	0.17	0.2
T→G	–	0.5	0.2	–	0.33	0.2
A→T	0.75	–	0.4	0.67	–	1.2
G→T	–	–	–	0.33	0.17	–
A→C	0.25	0.5	0.1	0.17	0.17	0.2
G→C	–	–	–	–	–	–
Sum: Transitions	2.5	2.75	17.7	3.67	3.5	3.8
Sum: Transversions	2.5	3.0	1.4	2.5	2.17	3.0
Mutation rate [% per position and PCR]	2.1	2.6	7.1	2.5	2.3	2.5

tations produced by a mutator strain are distributed randomly, a 1000 bp fragment or gene cloned into a high-copy plasmid (ColE1 origin of replication; 100 copies per cell; total length 4000 bp) could accumulate a maximum of 0.6 mutations per bp during 20 generations of growth.

The protocol for mutator strain passage and subsequent amplification of mutated plasmids is described in detail in the supplier's protocol. However, a condensed version is given here:

1. Transform XL1–Red according to the manufacturer's protocol and plate the transformants on LB media containing the appropriate antibiotic. Transformants become visible after 24–30 h. Often, longer incubation times are required.
2. For optimal random representation of mutations, pool all transformants from the agar plates by adding sterile LB broth (2 mL) to the plates and collecting the colonies with a sterile pipet.

3. Transfer the cell mass into a sterile cultivation tube, add antibiotic, and cultivate 1 h at 37 °C with shaking at 220 rpm.
4. The resulting culture represents ≈ 20 generations of growth and can be used directly to isolate the plasmid DNA with a standard miniprep protocol.
5. The DNA of interest can be amplified by transformation of a non-mutator strain (standard protocol) and cultivation on a 1–5 mL scale. The resulting culture can also be used directly for screening/selection purposes. Alternatively, the plasmid can be isolated by a standard procedure [21].
6. The mutator strain passage may be repeated for increasing the mutation frequency (see section 2.4, note 6).

Although the overall mutation frequency exhibited by a mutator strain is much lower than that of error-prone PCR, there are some advantages of this procedure. (1) A complete, supercoiled expression plasmid that has already been tested for suitability may be submitted to mutagenesis. (2) The loss of DNA material is minimal, compared to ligation and transformation of relaxed plasmid. (3) The mutation frequency (≈ 1 mutation per 2000 bp after one passage) can produce sufficient diversity for many optimization problems.

2.4 Troubleshooting

1. Do not mix 10X PCR buffer and Mn^{2+} solution, because this results in a precipitate that disables PCR amplification.
2. Routinely run a mutagenic PCR in parallel with the respective standard PCR and analyze a portion of these by agarose gel electrophoresis: If no PCR product is observed with the mutagenic PCR, try varying (mostly, increasing) the Mg^{2+} concentration in steps of 2 µL (= 2 mM).
3. A lack of product after mutagenic PCR may also be due to changed salt concentrations: try varying the annealing temperature in steps of 2 °C or use a gradient for approaching the correct annealing temperature.
4. With genes consisting of more than 1000 bp, it might be difficult to obtain amplification products after mutagenic PCR. If so, the template DNA may be divided into two or more fragments that are amplified separately.
5. Note that the ligation efficiencies for genes longer than 1000 bp with a standard expression vector (3000–3500 bp) are limited. Furthermore, the transformation efficiencies for relaxed plasmid DNA generated by ligation are drastically reduced, compared to supercoiled DNA. To solve the latter problem, competent cells of highest quality should be chosen for generating expressable mutant libraries.
6. Repeated mutator strain passage may promote recombination events, especially if the gene of interest exhibits toxic activity. Try chosing a lower number of passages.

References

1. E. C. Cox, *Annu. Rev. Genet.* **1976**, *10*, 135–156.
2. A. Greener, M. Callaghan, B. Jerpseth, *Methods Mol. Biol.* **1996**, *57*, 375–385.
3. D. Shortle, D. Nathans, *Proc. Natl. Acad. Sci. USA* **1978**, *75*, 2170–2174.
4. J. T. Kadonaja, J. R. Knowles, *Nucleic Acids Res.* **1985**, *13*, 1733–1745.
5. J. O. Deshler, *Gen. Anal. Techn. Appl.* **1992**, *9*, 103–106.
6. J. H. Spee, W. M. de Vos, O. P. Kuipers, *Nucleic Acids Res.* **1993**, *21*, 777–778.
7. D. W. Leung, E. Chen, D. V. Goeddel, *Technique* **1989**, *1*, 11–15.
8. R. C. Cadwell, G. F. Joyce, *PCR Methods Appl.* **1992**, *2*, 28–33.
9. J. P. Vartanian, M. Henry, S. Wain-Hobson, *Nucleic Acids Res.* **1996**, *24*, 2627–2631.
10. M. Zaccolo, D. M. Williams, D. M. Brown, E. Gherardi, *J. Mol. Biol.* **1996**, *255*, 589–603.
11. M. Suzuki, F. C. Christians, B. Kim, A. Skandalis, M. E. Black, L. A. Loeb, *Molec. Div.* **1996**, *2*, 111–118.
12. M. S. Horwitz, L. A. Loeb, *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 7405–7409.
13. L. A. Loeb, *Adv. Pharmacol.* **1996**, *35*, 321–347.
14. L. J. Jensen, K. Andersen, A. Svendsen, T. Kretzschmar, *Nucleic Acids Res.* **1998**, *26*, 697–702.
15. K. R. Tindall, T. A. Kunkel, *Biochemistry* **1988**, *27*, 6008–6013.
16. J. Cline, J. C. Braman, H. H. Hogrefe, *Nucleic Acids Res.* **1996**, *24*, 3546–3551.
17. K. A. Eckert, T. A. Kunkel, *PCR Methods Appl.* **1991**, *1*, 17–24.
18. R. A. Beckman, A. S. Mildvan, L. A. Loeb, *Biochemistry* **1985**, *24*, 5810–5817.
19. D. H. Gelfand, T. H. White, in: *PCR Protocols: A Guide to Methods and Applications*, (M. A. Innis, D. H. Gelfand, J. J. Sninsky and T. J. White, eds.), Academic Press, San Diego, 1990 pp. 129–141.
20. R. C. Cadwell, G. Joyce, *PCR Methods Appl.* **1994**, *4*, S136–S140.
21. J. Sambrook, E. F. Fritsch, T. Maniatis, *Molecular Cloning: A Laboratory Manual*, 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, **1989**.
22. A. Greener, M. Callaghan, B. Jerpseth, *Mol. Biotechnol.* **1997**, *7*, 189–195.

3 DNA Shuffling

Hikaru Suenaga, Masatoshi Goto, and Kensuke Furukawa

3.1 Introduction

DNA shuffling mimics the process of natural evolution, in which the immense diversity of all life forms has been created. It generates diversity by recombination, combining useful mutations from individual genes (Fig. 3.1). In DNA shuffling, the diversity implies libraries of related, chimeric genes. The libraries can be generated by random fragmentation of a pool of related genes, followed by reassembly of the fragments by self-priming PCR. This process causes crossovers between homologous sequences, due to template switching (Fig. 3.1).

DNA shuffling consists of 5 steps (Fig. 3.2): (1) preparation of parent genes, (2) random fragmentation of parent genes with DNase I, (3) recovery of DNA fragments in specific molecular size ranges, (4) reassembly of these fragments by primerless PCR (self-priming PCR), and (5) amplification of reassembled products

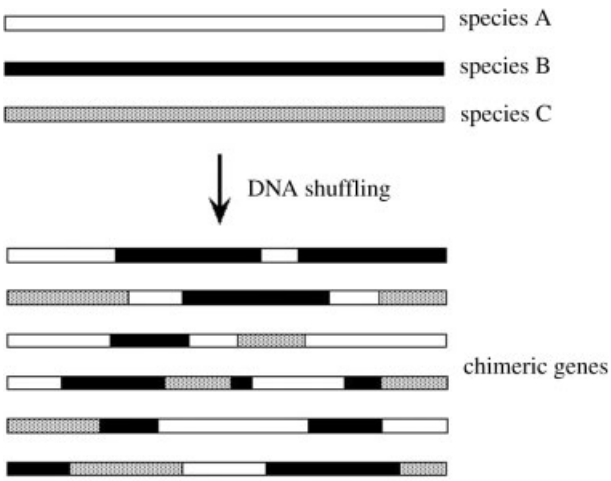


Fig. 3.1. Evolution of genes by DNA shuffling. Homologous recombination of evolutionarily related genes creates a library of chimeric genes. Some genes from related species are 'shuffled' to create an even larger pool of novel genes.

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst
Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-30799-0

by conventional PCR with primers. PCR products are cloned into a vector for expression, followed by screening or selection. Multiple cycles of DNA shuffling can be used to evolve the desired properties.

In DNA shuffling starting from a single gene as the parent template, diversity originates from random point mutations, due to the limited fidelity of polymerases

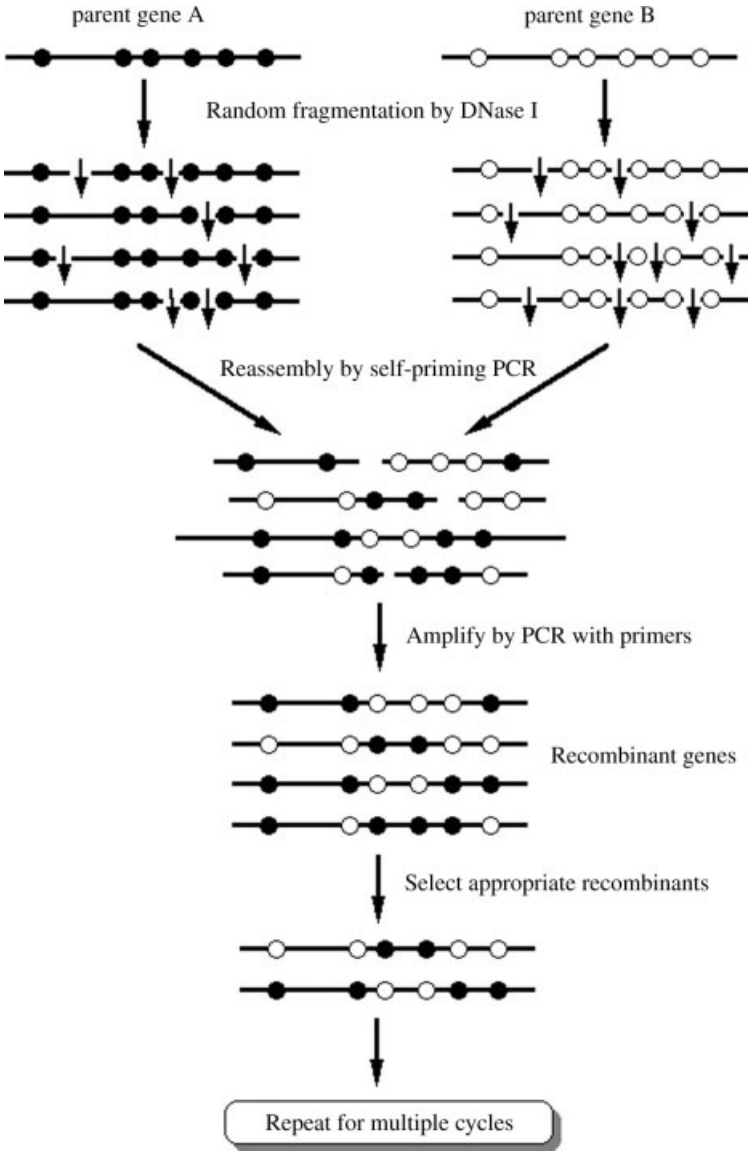


Fig. 3.2. Schematic representation of DNA shuffling using 2 parental genes.

used in PCR [1,2]. After screening, improved clones are used as template sequences for the next process of DNA shuffling to recombine useful mutations in additive or synergistic ways. Although these point mutations may provide useful diversity, the high mutation rate decreases the frequency of active clones [3]. Zhao and Arnold developed a high-fidelity DNA recombination protocol using Mn^{2+} instead of Mg^{2+} and proofreading DNA polymerases (*Pfu* and *Pwo*) instead of *Taq* polymerase in the step of random fragmentation by DNase I and PCR, respectively [4, 5].

A recent adaptation, called family shuffling, allows more than 2 genes, e. g., genes from different species, to be used as the parental sequences [6]. The genes in a library created by single gene DNA shuffling differ by only a few point mutations [1–5]. In contrast, the block-exchange nature of family shuffling creates chimeras that differ in many positions. Such related genes provide a greater functional diversity than conventional DNA shuffling using a single gene [6–9]. It is difficult to generalize the prerequisite homologies between parent genes for successful family shuffling. According to our experience, homologies of at least 80% and 60% are necessary in DNA family shuffling using 2 genes and 3 or more genes, respectively. In this chapter, we describe a procedure for DNA shuffling that uses 2 or more parental genes and exhibits high fidelity.

3.2 Materials

3.2.1 For Preparation of Parental Genes

- dH_2O (sterilized deionized or distilled water)
- Restriction endonuclease
- 10× restriction endonuclease buffer
- TE buffer:
  10 mM Tris-Cl (pH 8.0)
  1 mM EDTA (ethylenediaminetetraacetic acid) (pH 8.0)
- Gel extraction kit

  Many companies supply these kits, for example, the QIAquick Gel Extraction Kit from Qiagen or the TaKaRa Recochip from Takara Bio Inc.

3.2.2 For Random Fragmentation by DNase I

- dH_2O
- 10× digestion buffer:
  1 M Tris-Cl (pH 5.0)
  10 mM $MnCl_2$
- DNase I (Takara Bio Inc.)
- 150 mM NaCl

- 0.5 M EDTA (pH 8.0)
- 3 M sodium acetate (pH 5.2)
- 99% ethanol
- 70% ethanol
- TE buffer

3.2.3 For Collection of DNA Fragments in Specific Molecular Size Ranges

- Low-melting-point agarose gel (Bethesda Research Laboratories)
- 10 bp DNA step ladder (Promega)
- DE81 ion-exchange paper (Whatman) or TaKaRa Recochip (Takara Bio Inc.)
- 1 M NaCl
- Phenol/chloroform/isoamyl alcohol mixture (25:24:1)

  This mixture consists of equal parts of equilibrated phenol and chloroform:isoamyl alcohol (24:1). Neither chloroform nor isoamyl alcohol requires treatment before use. The mixture may be stored under 100 mM Tris-Cl (pH 8.0) in a light-tight bottle at 4 °C for up to 1 month.

  Alternatively, Sigma supplies phenol:chloroform:isoamyl alcohol 25:24:1 saturated with 10 mM Tris, pH8.0, 1 mM EDTA for use in molecular biology.

3.2.4 For Reassembly of These Fragments by Primerless PCR

- 2× PCR premixed solution (final concentrations):
  0.4 mM each dNTP (deoxynucleoside triphosphate)
  5-fold diluted *Pfu* DNA polymerase buffer
  0.125 U μL^{-1} native or cloned *Pfu* DNA polymerase

Prepare an adequate volume of premixed solution each time you start an experiment. Stored solutions must not be used.

Example preparation for one sample (10 µL):

0.4 µL of 10 mM each dNTP (PCR Nucleotide Mix, Promega)
2 µL of *Pfu* DNA polymerase buffer (Stratagene)
0.5 µL of native or cloned *Pfu* DNA polymerase (2.5 U μL^{-1}, Stratagene)
5.5 µL of dH_2O

3.2.5 For Amplification of Reassembled Products by Conventional PCR with Primers

- dH_2O
- 10 mM each dNTP
- *Taq* DNA polymerase (Promega)

- *Pfu* DNA polymerase (Stratagene)
- 10× *Taq* DNA polymerase buffer with $MgCl_2$
- A set of primers (forward and reverse)
- PCR purification kit
  Many companies supply these kits, for example, the QIAquick PCR Purification Kit (Qiagen) or Montage^{TM} PCR Centrifugal Filter Devices (Millipore).

3.3 Protocol

3.3.1 Preparation of Parental Genes

1. Digest plasmids (5 µg) containing parent gene with appropriate restriction endonucleases.
2. Electrophorese and recover DNA fragments of parent gene from agarose gels using gel extraction kit (see Section 3.2.1).
3. Dissolve the DNA fragments in TE buffer and measure the concentration of each parental DNA in solution.
4. Mix in equal proportions for a total of 2 µg.

The preparation of parent genes by PCR is effective when it is difficult to excise a target gene region from a plasmid because of the absence of a proper restriction site. In this case, it is important to purify the PCR products completely by using a PCR purification kit.

3.3.2 Random Fragmentation by DNase I

In this step, prepare 4 tubes to digest DNA for various incubation times (2, 4, 6, and 8 min).

1. Add 10 µL of 10× digestion buffer (see Section 3.2.2) and bring the volume to 98 µL with water.
2. Equilibrate mixture at 15 °C for 10 min.
3. Add 2 µL of DNase I (0.15 U) freshly diluted to 0.075 U μL^{-1} in 150 mM NaCl and perform digestion at 15 °C.
4. After 2, 4, 6, and 8 min of incubation, add 4 µL of 0.5 M EDTA solution to each of the 4 tubes. Next, heat them at 80 °C for 10 min to stop the reaction completely.
5. Mix the contents of all 4 tubes. Subsequently, add 40 µL of 3 M sodium acetate and 800 µL of 99% cold ethanol.
6. After incubation on ice for 60 min (or −20 °C for 30 min), centrifuge in a microcentrifuge at maximum speed for 10–20 min.
7. Carefully aspirate the ethanol solution with a micropipette. Remove the supernatant as completely as possible.

8. Rinse the pellet by adding 500 µL of 70% ethanol. At this point it is not necessary to centrifuge.
9. Carefully aspirate all the ethanol solution with a micropipette. Be careful not to disturb the pellet, which may or may not be visible.
10. Dry the pellet in a vacuum centrifuge.
11. Dissolve the DNA in TE buffer. All the DNA is used in the next step.

3.3.3 Collection of DNA Fragments in Specific Molecular Size Ranges

The procedures in this section are represented in Fig. 3.4.

1. Separate fragments by electrophoresis on 2% low-melting-point agarose gel.

To avoid overload, the sample of DNA with the gel-loading dye should be loaded into a wide slot of the gel. The DNA molecular weight marker is usually run in the two outside wells of the gel. A large gel (> 10 cm long) should be prepared and be run slowly (< 1 V cm^{-1}) for 10–12 h to obtain maximum resolution of the DNA fragments. Electrophoresis at high voltage (10 V cm^{-1}) shortens the running time (1–1.5 h); however, it is likely to decrease the resolution of the DNA fragments. DNA of a given size runs slightly faster through gels cast with low-melting-point agarose than through conventional agarose gels.

2. Locate the bands using a handheld long-wavelength ultraviolet lamp to minimize radiation damage to the DNA.
3. Using a sharp razor blade, remove the agarose gel containing DNA fragments above 50 bp. Next, make an incision in the gel directly in front of the leading edge of the band of interest and about 2 mm wider than the band on each side.
4. Cut a piece of DE81 ion-exchange paper the same width as the incision and slightly longer (2 mm) than the gel's thickness.
5. Using blunt-ended forceps, soak the paper in the running buffer and insert it into the slit, being careful not to trap air bubbles.
6. Resume electrophoresis (5–10 V cm^{-1}) until the bands of DNA have migrated onto the paper. Follow the progress of the electrophoresis with a handheld long-wavelength ultraviolet lamp.
7. When all of the target DNA is trapped on the paper, remove it from the gel. Using forceps, transfer the paper to a 1.5-mL microtube.
8. Add enough volume of 1 M NaCl to cover the paper completely and incubate for 10 min at room temperature.
9. Transfer the fluid to a new microtube. Repeat the elution (step 8 above) 3 times and combine all fluids.
10. Extract the eluant twice with 2 volumes of phenol/chloroform/isoamyl alcohol mixture (25 : 24 : 1).
11. Add 0.1 volume of 3 M sodium acetate and 2 volumes of cold 99% ethanol to the eluant.
12. Perform ethanol precipitation (see Section 3.3.2, steps 6–10).

13. Dissolve DNA in 15 µL of TE buffer and, to check for sufficient DNA concentration, electrophorese an aliquot of the DNA solution.

If desired, a TaKaRa Redochip can be used instead of the DE81 ion-exchange paper. In this case, the elution, extraction, and ethanol precipitation steps (steps 8–12 above) can be omitted.

3.3.4 Reassembly of These Fragments by Primerless PCR

1. Combine 10 µL of purified DNA (100–200 ng) and 10 µL of 2× PCR premixed solution (see Section 3.2.4).
2. Run the assembly reaction using the following thermocycle program:
   a) 5 min at 94 °C
   b) 1 min at 94 °C
   c) 1 min at 52 °C (if possible: 50–55 °C in thermogradient mode)
   d) 1 min + 5 sec per cycle at 72 °C*
      Perform 45 cycles of steps b–d.
   e) 10 min at 72 °C

   * If the thermocycler has no time increment function, use the following conditions for the extension time: 1 min + 15 sec for each of 15 cycles.

It is important to optimize each thermocycler program. In general, small DNA fragments may require a lower annealing temperature and more cycles for assembly. and large parent genes may need a longer extension time. Using a thermocycler that is equipped with a thermogradient mode is effective for determining the optimal annealing temperature.

3. Run a small aliquot of the amplified products on an agarose gel to evaluate if sufficient amplification has occurred. A smeared band can be seen (Fig. 3.3).

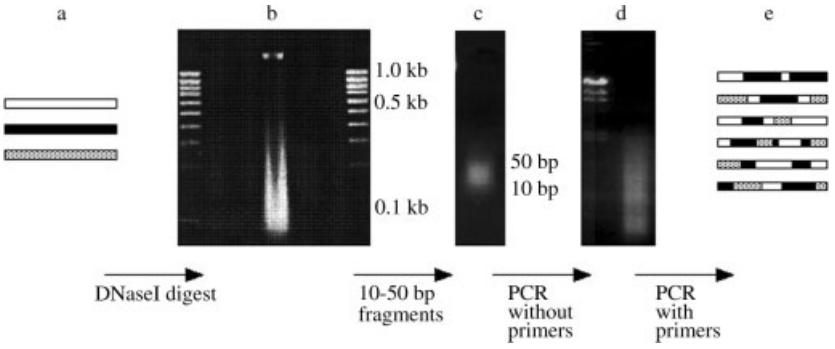


Fig. 3.3. Agarose gel electrophoresis profiles in the process of DNA shuffling. (a) Parental DNA fragments of interest are prepared. (b) Template DNAs are digested with DNase I. (c) DNA fragments of 10–50 bp are purified from agarose gel. (d) Purified fragments are reassembled into a full-length DNA in the absence of primers. (e) A single PCR product of the correct size is typically obtained by PCR with primers.

3.3.5 Amplification of Reassembled Products by Conventional PCR with Primers

1. Carry out a standard PCR using the corresponding primers and the primerless PCR product as template. PCR conditions (50 µL final volume) are: 1 µL of dilute (1:1000, 1:500, 1:100, 1:50, and 1:10) primerless PCR products, 50 pmol of each primer, 1× *Taq* DNA polymerase buffer containing $MgCl_2$, 0.2 mM of each dNTP and 1.25 U *Taq*/*Pfu* (1:1) mixture. The PCR program (25 cycles) is the same as for primerless PCR, except that the extension time is constant:
   a. 5 min at 94 °C
   b. 1 min at 94 °C
   c. 1 min at 52 °C (50–55 °C in thermogradient mode)
   d. 1 min at 72 °C
   Perform 25 cycles of steps b–d.
   e. 10 min at 72 °C
2. Confirm by agarose gel electrophoresis that a single band of the correct size appears (Fig. 3.4).

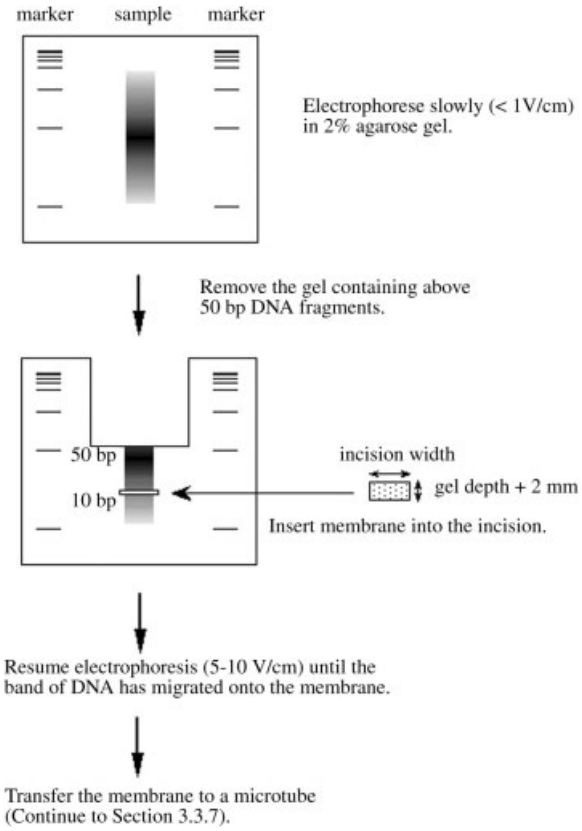


Fig. 3.4. Procedure for recovery of small DNA fragments with DEAE ion-exchange paper.

3. Purify the reaction products by using a PCR purification kit or by ethanol precipitation (see Section 3.3.2 steps 5–10), digest with appropriate restriction endonucleases, and ligate into the cloning vector.

3.4 Troubleshooting

3.4.1 Insufficient DNase I Fragmentation

The concentration of Mn^{2+} is not optimal.

- The reaction rate of DNase I is generally affected by the Mn^{2+} concentration. Increase or decrease the concentration.

3.4.2 Little or No Product of Primerless PCR

1. The concentration of template DNA is insufficient.
   - Increase the amount of starting template used in PCR (> 100 ng).
2. The DNA is not sufficiently purified.
   - Repeat the phenol/chloroform/isoamylalcohol extraction step (Section 3.3.3 steps 3–5). Alternatively, use a DNA purification kit, which also provides good results.
3. The concentration of *Pfu* DNA polymerase is too low.
   - Increase the amount of *Pfu* DNA polymerase in 0.5-unit steps.
4. The annealing temperature is incorrect.
   - Decrease the annealing temperature in 2 °C steps.
5. The number of cycles is insufficient.
   - Repeat the step of primerless PCR (Section 3.3.4 steps 1–2) by using 10 µL of the (insufficient) first primerless PCR product. This is more effective than increasing of the number of PCR cycles.
6. The homology of the parental genes is too low.
   - Low homologies of DNA fragments reduce the incidence of assembly in the process of primerless PCR. To increase the frequency of association, use larger DNA fragments. Recover 50- to 100-bp DNA fragments in the DNA-collection step (Section 3.3.3 step 2). However, the use of larger DNA fragments in primerless PCR decreases the diversity of recombination.

3.4.3 Little or No Product of PCR with Primers

1. The concentration of template DNA is insufficient.
   - Increase the amount of template used in PCR.
2. Extension time is too short.
   - Increase the extension time in increments of 1 min.

3. The concentration of DNA polymerase is too low.
   - Increase the amount of DNA polymerase mixture in 0.5-unit steps.
4. The number of cycles is insufficient.
   - Increase the number of cycles in steps of 5 cycles.
5. The annealing temperature is incorrect.
   - Decrease the annealing temperature in 2 °C steps.
6. The primer concentration is too low.
   - Increase the primer concentration up to 1.0 μM.

3.4.4 The Product of PCR with Primers is Multi-banded

1. The annealing condition is incorrect.
   - Increase the annealing temperature in 2 °C steps and shorten the annealing time to 30 s.
2. The Mg^{2+} concentration is not optimal.
   - Perform PCR with different final concentrations of Mg^{2+}, from 1.5 to 5 mM in 0.5 mM steps.
3. Primers anneal incorrectly with complicated templates.
   - Difficulties in determining the optimal annealing temperature can often be overcome by performing touchdown PCR [10] or nested PCR [11].

3.5 Amplification Examples

DNA shuffling is a powerful tool for directed evolution of gene products toward desired properties such as enhanced activity [12–15], improved protein folding [16–19], and altered substrate specificity [20–23]. We have shown that DNA shuffling is extremely useful for creating optimized enzymes, e. g., dioxygenases for the efficient degradation of environmental pollutants [24, 25].

Naturally occurring biphenyl dioxygenase is involved in the initial oxygenation and subsequent degradation of biphenyl and polychlorinated biphenyls (PCBs), a family of xenobiotic compounds that are environmental pollutants [26–28]. The biphenyl dioxygenases of strains *Pseudomonas pseudoalcaligenes* KF707 and *Burkholderia cepacia* LB400 exhibit distinct differences in substrate specificity and regiospecificity of oxygen insertion, and hence, in their ability to biodegrade PCBs, despite the fact that these two enzymes are nearly identical in amino acid sequence [29–31]. Biphenyl dioxygenases are multicomponent enzymes in which a large subunit, encoded by the *bphA1* gene, is significantly responsible for substrate recognition [32, 33]. Using the process of DNA shuffling of two *bphA1*, a number of evolved biphenyl dioxygenases were created (Fig. 3.5). Some of these evolved enzymes exhibited enhanced degradation capacity, not only for PCB and related biphenyl compounds, but also for single aromatic hydrocarbons such as ben-

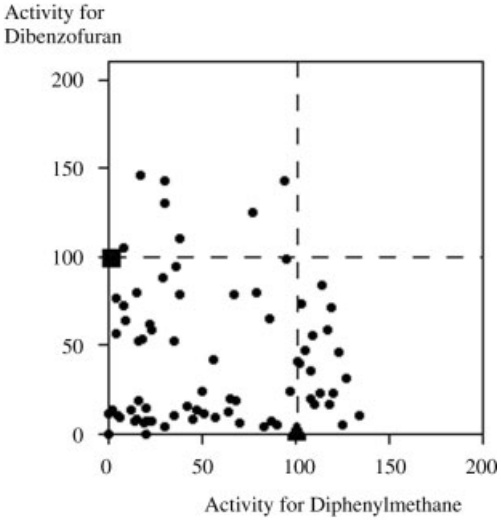


Fig. 3.5. Degradation activities of various chimeric biphenyl dioxygenases against dibenzofuran and diphenylmethane. KF707 biphenyl dioxygenase exhibits degradation activity for only diphenylmethane, as shown by the large triangle on the x axis (set to 100) and that LB400 biphenyl dioxygenase exhibits activity for only dibenzofuran, as shown by the large square on the y axis (set to 100). The activities of the evolved enzymes are shown by the circles.

zene and toluene, which are poor substrates for both parent biphenyl dioxygenases [20, 24].

Recently, Zhang and coworkers reported whole-genome shuffling. This method combines the advantage of DNA family shuffling with the recombination of entire genomes used in conventional breeding [34]. This approach provides a nonrecombinant alternative for rapidly improving organisms.

One major advantage of *in-vitro* DNA shuffling of enzymes over the structural remodeling is that little prior information is required. Since the structures of a large number of proteins are being solved by X-ray analyses, structural information might be used properly to guide future strategies of DNA shuffling.

References

1. W.P.C. Stemmer, *Proc. Natl. Acad. Sci. USA*, **1994**, *91*, 10747–10751.
2. W.P.C. Stemmer, *Nature*, **1994**, *370*, 389–391.
3. J.C. Moore, F.H. Arnold, *Nature Biotech.*, **1996**, *14*, 458–467.
4. H. Zhao, F.H. Arnold. *Nucleic Acids Res.*, **1997**, *25*, 1307–1308.
5. H. Zhao, F.H. Arnold. *Proc. Natl. Acad. Sci. USA*, **1997**, *94*, 7997–8000.
6. A. Crameri, S.-A. Raillard, E. Bermudez, W.P.C. Stemmer, *Nature*, **1998**, *391*, 288–291.
7. F.C. Christians, L. Scapozza, A. C. Crameri, G. Folkers, W. P. C. Stemmer, *Nature Biotechnol.*, **1999**, *17*, 259–264.
8. C.C.J. Chang, T.T. Chen, B.W. Cox, G.N. Dawes, W.P.C. Stemmer, J. Punnonen, P.A. Patten, *Nature Biotechnol.*, **1999**, *17*, 793–797.
9. J.E. Ness, M. Welch, L. Giver, M. Bueno, J.R. Cherry, T.V. Borchert, W.P.C. Stemmer, J. Minshull. *Nature Biotechnol.*, **1999**, *17*, 893–896.

10. R.H. Don, P.T. Cox, B.J. Wainwright, K. Baker, J.S. Mattick. *Nucleic Acids Res*., **1991**, *19*, 4008.
11. C.W. Dieffenbach, G.S. Dveksler. *PCR Primer: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, **1995**.
12. F. Buchholz, P.O. Angrand, A.F. Stewart. *Nature Biotechnol*., **1998**, *16*, 657–662.
13. S.R. Leong, J.C. Chang, P. Org, G. Dawes, W.P. Stemmer, J. Punnonen. *Proc. Natl. Acad. Sci. USA*, **2003**, *100*, 1163–1168.
14. W.M. Coco, L.P. Encell, W.E. Levinson, M.J. Crist, A.K. Loomis, L.L. Licato, J.J. Arensdorf, N. Sica, P.T. Pienkos, D.J. Moticello. *Nature Biotechnol*., **2002**, *20*, 1246–1250.
15. P.R. Salamone, I.H. Kavakli, C.J. Slattery, T.W. Okita. *Proc. Natl. Acad. Sci. USA*, **2002**, *99*, 1070–1075.
16. A. Crameri, E.A. Whitehorn, E. Tate, W.P.C. Stemmer. *Nature Biotechnol*., **1996**, *14*, 315–319
17. K. Proba, A. Worn, A. Honegger, A. Pluckthun. *J. Mol. Biol*., **1998**, *275*, 245–253.
18. P. Martineau, P. Jones, G. Winter. *J. Mol. Biol*., **1998**, *280*, 117–127.
19. J.D. Wang, C. Herman, K.A. Tripton, C.A. Gross, J.S. Weissman. *Cell*, **2002**, *111*, 1027–1039.
20. T. Kumamaru, H. Suenaga, M. Mitsuoka, T. Watanabe, K. Furukawa. *Nature Biotechnol*., **1998**, *16*, 663–666.
21. T. Yano, S. Oue, H. Kagamiyama. *Proc. Natl. Acad. Sci. USA*, **1998**, *95*, 5511–5515.
22. D.R. Liu, T.J. Magliery, M. Pastrnak, P.G. Schultz. *Proc. Natl. Acad. Sci. USA*, **1997**, *94*, 10092–10097.
23. M.D. Van Kampen, N. Dekker, M.R. Egmond, H.M. Verheij. *Biochemistry*, **1998**, *37*, 3459–3466.
24. H. Suenaga, M. Mitsuoka, Y. Ura, T. Watanabe, *J. Bacteriol*., **2001**, *183*, 5441–5444.
25. K. Furukawa, *Curr. Opin. Biotechnol*., **2000**, *11*, 244-249.
26. K. Furukawa, *Biodegradation and Detoxification of Environmental Pollutants*. CRC Press, Boca Raton, FL, **1982**, 33–57.
27. K. Furukawa, *Biodegradation*, **1994**, *5*, 289–300.
28. R. Unterman, *Bioremediation*, Cambridge University Press, New York, **1996**, 209–253.
29. D.T. Gibson, D.L. Cruden, J.D. Haddock, G.J. Zylstra, J.M. Brand, *J. Bacteriol*., **1993**, *175*, 4561–4564.
30. B.D. Erickson, F.J. Mondello, *J. Bacteriol*., **1992**, *174*, 2903–2912.
31. K. Taira, J. Hirose, S. Hayashida, K. Furukawa, *J. Biol. Chem*. **1992**, *267*, 4844–4853.
32. N. Kimura, A. Nishi, M. Goto, *J. Bacteriol*., **1997**, *179*, 3936–3943.
33. H. Suenaga, A. Nishi, T. Watanabe, M. Sakai, K. Furukawa, *J. Biosci. Bioeng.* **1999**, *87*, 430–435.
34. Y.-X. Zhang, K. Perry, V.A. Vinci, K. Powell, W.P.C. Stemmer, S.B.D. Cardayre, *Nature*, **2002**, *415*, 644–646.

4 DNA Recombination Using StEP

Milena Ninkovic

4.1 Introduction

In vitro recombination holds a central role in directed evolution, with many different strategies having been developed (for review see [9]). This article tries to illustrate one of the most popular *in vitro* DNA recombination methods, the staggered extension process (StEP). Nature uses recombination to speed up evolution, which would otherwise depend on the sequential accumulation of (favorable) point mutations. The mechanism of natural recombination is straightforward: search for homology, break homologous chromosomes, exchange strands, and ligate the newly recombined DNA strands. Since recombination can occur even with 'naked' DNA [10], *in vitro* recombination has been successfully employed to produce variant proteins with a wide range of modified properties (for review, see [12]) or even to evolve whole operons [3].

The StEP method, developed by Zhao et al. [14], is a technically simple method, which generates full-length recombined genes in the presence of template(s). Using this method, it is possible, during only one PCR, to create a library of recombined DNA sequences. In practice, the StEP method consists of: (i) very short annealing/extension steps in which the primer extension is limited and (ii) a subsequent denaturation step in which the extension is interrupted by heat denaturation. In each cycle the growing fragments can anneal to different templates, based on sequence complementarity, and are extended further. Repeating these cycles of partial extension and denaturation finally creates a library of recombined full-length sequences. The procedure is illustrated in Figure 4.1. The aim of this method is to achieve higher crossover frequency through more template switches. One way to achieve this goal is to use special DNA polymerases. Usually, these enzymes are very fast. Even very brief extension periods provide enough time for these enzymes to extend primers by hundreds of nucleotides (nt), e. g., at 70 °C, more than 60 nt s^{-1}, at 55 °C approximately 24 nt s^{-1}, at 37 °C approximately 1.5 nt s^{-1}, and at 22 °C approximately 0.25 nt s^{-1} [5]. However, not all polymerases are equally fast. Polymerases with proofreading activity, e. g., *Pfu* and *Vent* DNA polymerases, are slower and therefore may be more suited to enhancing crossover frequency in a StEP process. However, the use of proofreading polymerases also minimizes the rate of associated point mutations. Accompanying point mutations might sometimes be desired but

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst
Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-30799-0

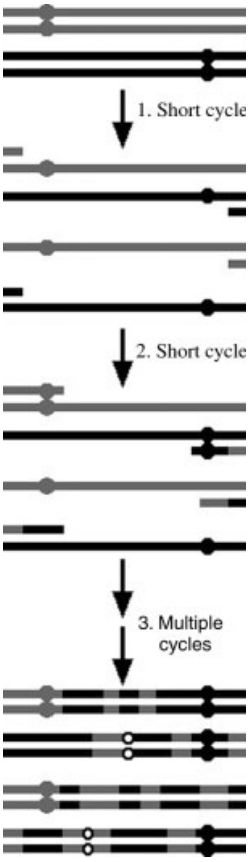


Fig. 4.1. Schematic representation of the StEP process using two parental DNA sequences. (1) Denatured template DNAs are primed with defined primers. (2) The partially extended primers produced by very brief annealing/extension randomly reanneal to different parent sequences (template switching). (3) Novel recombinants are created through multiple cycles of annealing/extension and strand switching. In principle, StEP is also an error-prone amplification process that introduces additional point mutations (white circles).

can also cause problems, e. g., in the recombination of large genes [8] in which the fraction of genes generating active proteins rapidly decreases with the number of acquired codon replacements [11].

4.2 Materials

4.2.1 StEP PCR

- *Taq* DNA polymerase (Promega, Stratagene, Roche)
- *Vent* DNA polymerase (New England Biolabs)
- Corresponding 10× polymerase buffer with $MgCl_2$ (for *Taq* DNA polymerase) or $MgSO_4$ (for *Vent* DNA polymerase)

- dNTP mix containing 2 mM of each dNTP (Pharmacia)
- Set of primers (forward and reverse)
- dH_2O (double-distilled water)

4.2.2 Purification of an Appropriate DNA Fragment

- Phenol/chloroform/isoamyl alcohol mixture (25:24:1; Sigma)
- 2.5 M sodium acetate, pH 5.2
- 95% ethanol
- 70% ethanol
- PCR purification kit or gel extraction kit (QIAquick PCR Purification Kit, QIAquick Gel Extraction Kit, both from Qiagen, or NucleoSpin® Extract Kit from Macherey-Nagel)

4.2.3 Equipment

- Thermocycler (e. g., UNO II Thermocycler, Biometra, Germany)
- Thin-wall PCR tubes

4.3 Protocol

1. Mix equimolar amounts of different DNA templates. Templates can be plasmids carrying the target sequence, (purified) PCR products, or DNA excised by restriction endonucleases.
2. To 20 ng of total template DNA, add 5 µL of appropriate 10× DNA polymerase buffer, 5 µL of dNTP mix, 10–100 pmol of each primer, 2.5 U *Taq* DNA polymerase or 1–2 U *Vent* DNA polymerase; add dH_2O to a final volume of 50 µL.
3. Run the reaction using the following thermocycler program:
   - 5 min at 95 °C
   - 20–100 cycles (depending on thermocycler and template used) of:
     - 30 s at 95 °C
     - 1 s at 55 °C
4. Check samples of the reaction by agarose gel electrophoresis.
5. If there is only a single band at the correct apparent molecular weight, extract the PCR reaction mixture twice with 1 volume of phenol/chloroform/isoamyl alcohol mixture. Otherwise, see step 7.
6. Perform an ethanol precipitation by adding 1/10 volume of 2.5 M sodium acetate and 2–3 volumes of 95% ice-cold ethanol. Hold the solution at −80 °C for 30 min or at −20 °C for 2 h. After centrifugation in a microcentrifuge at 4 °C and maximum speed for 30 min, carefully remove the supernatant with a mi-

cropipette. Rinse the pellet by adding 500 µL of 70% ethanol; then centrifuge for 5 min at room temperature and maximum speed. Carefully remove all ethanol solution and briefly dry the pellet, e. g., in a SpeedVac. Dissolve the DNA in dH_2O.

Alternatively, the PCR amplification product may be purified using a standard PCR purification kit.

Go to step 8.

7. If analytical electrophoresis reveals more than one band in the agarose gel (step 4), purify the desired DNA fragment by preparative agarose gel electrophoresis and subsequent elution, using a gel extraction kit.
8. With the purified DNA, perform a digestion with the appropriate restriction endonucleases and ligate the cleavage product into an appropriate cloning vector.

4.4 Technical Tips

After optimal conditions for recombination/PCR amplification have been found, it is highly recommended to always use the same thermocycler (ultrafast thermocyclers may be advantageous) and thin-wall PCR tubes for maximal reproducibility.

4.4.1 Problem: Little or No PCR Product (Full-length Product) after PCR

- For optimization of the amplification reaction, repeat the amplification with up to 100 cycles. If there is still little or no product, you can use a second (normal) PCR seeded by an aliquot of the first PCR to increase the yield of the StEP reaction.
- It is also important to optimize the length of the annealing/extension step. You should start with a shorter annealing/extension time. Because of the very fast polymerase activity, very often full-length product can be achieved after only 10–20 cycles, but with lower crossover frequency.

4.4.2 Problem: High Background Levels of DNA after PCR

Parental template DNA (plasmid DNA or short DNA amplified by PCR sequence or excised by restriction endonucleases) can be a problematic contaminant leading to the production of either wild-type clones or other ‘false’ recombination products. If this happens, remove the parent DNA. There are several ways of achieving this:

- If you use plasmid DNA as a template, purify the PCR product by agarose gel electrophoresis and subsequent elution using a DNA extraction kit.

- If you use short DNA fragments as templates that are indistinguishable in size from the product, conventional physical separation, e. g., agarose gel electrophoresis, cannot be employed. In this case, enzymatic degradation of the parent DNA, e. g., by *Dpn I* digestion, may be necessary to reduce the background. The *Dpn I* restriction enzyme cleaves at the recognition site GATC only when it is fully methylated (N^6-methylation of adenine) on both strands, while leaving hemi-methylated or non-methylated DNA intact [7]. Therefore, the *Dpn I* enzyme specifically removes parent DNA isolated from a dam^+ *E. coli* strain, such as XL1-Blue or DH5α. Slightly more demanding is the use of biotinylated primers. If short DNA (e. g., PCR products) are used as templates, biotinylated primers can be used to enrich recombined DNA after StEP PCR. Alternatively, template DNA biotinylated on both strands can be used.

4.5 StEP in Directed Evolution

The StEP recombination method was first used to improve the thermostability of subtilisin E, a protease produced by the mesophile *Bacillus subtilis*. Five thermostable mutants were recombined and yielded an enzyme whose half-life at 65 °C is 50 times greater than that of wild-type subtilisin E [14]. The StEP method was also used for directed evolution in an effort to modify the regioselectivity of α-galactosidase AgaB of *Bacillus stearothermophilus* [4] and to expand the substrate specificity of biphenyl dioxygenase [2]. Although for some applications, point mutations introduced by using *Taq* DNA polymerase may provide more diversity, for some applications mutations can be problematic. For example, much lower mutation rates are needed for the *in vitro* evolution of long genes or whole operons. Structure-function studies of evolutionarily related sequences have revealed that an increased number of mutations is often correlated with inactivation of the protein(s) [1, 11]. When an alternative protocol that relies on the proofreading *Vent* DNA polymerase was used, the mutation rate was only 0.02% [8], which is at least one-third the mutation rate obtained with previous protocols [13, 14]. Comparison of the maximal extension rate of 1000 nt min^{-1} and a processivity of about 7 nucleotides per initiation event for *Vent* DNA polymerase with 4000 nt min^{-1} and 40 nucleotides per initiation event for *Taq* DNA polymerase [6] also shows that *Vent* DNA polymerase should be preferred when high crossover frequency due to more template switches is desired.

The efficiency of StEP recombination is similar to that of other *in vitro* recombination methods, but an advantage of this method lies in the fact that the reaction can be carried out in a single test tube. Template DNA can be double-stranded as well as single-stranded, and there is no significant restriction to the number of parent templates to be recombined.

References

1. Bowie, J.U., J. Reidhaar-Olson, W.A. Lim, and R.T. Sauer **1990**. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science 247*: 1306–1310.
2. Bruehlmann, F. and W. Chen **1999**. Tuning biphenyl dioxygenase for extended substrate specificity. *Biotechnol. Bioeng. 63*: 544–551.
3. Crameri, A., G. Dawes, E. Rodriguez, S. Silver and W.P.C. Stemmer **1997**. Molecular evolution of an arsenate detoxification pathway by DNA shuffling. *Nature Biotechnol. 15*: 436–438.
4. Dion, M., A. Nisole, P. Spangenberg, C. André, A. Glottin-Fleury, R. Mattes, C. Tellier and C. Rabiller **2001**. Modulation of the regioselectivity of a *Bacillus* α-galactosidase by directed evolution. *Glycoconjugate J. 18*: 215–223.
5. Innis, M.A., K.B. Myambo, D.H. Gelfand and M.A. Brow **1988**. DNA sequencing with *Thermus aquaticus* DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. *Proc. Natl. Acad. Sci. USA 85*: 9436–9440.
6. Kong, H., R.B. Kucera and W.E. Jack **1993**. Characterization of a DNA polymerase from the hyperthermophile archaea *Thermococcus litoralis*: *Vent* DNA polymerase, steady state kinetics, thermal stability, processivity, strand displacement, and exonuclease activities. *J. Biol. Chem. 268*: 1965–1975.
7. Nelson, M. and M. McClelland **1992**. Use of DNA methyltransferase/endonuclease enzyme combinations for megabase mapping of chromosomes. *Meth. Enzymol. 216*: 279–303.
8. Ninkovic, M., R. Dietrich, G. Aral and A. Schwienhorst **2001**. High fidelity in vitro recombination using a proofreading polymerase. *BioTechniques 30*: 530–536.
9. Ninkovic, M., R. Dietrich and A. Schwienhorst **2001**. Advances in DNA recombination technology. *Biotech. News Int. 6*: 14–15.
10. Stemmer, W.P.C. **1994**. Rapid evolution of a protein in vitro by DNA shuffling. *Nature 370*: 389–391.
11. Suzuki, M., F.C. Christians, B. Kim, A. Skandalis, M. Black and L.A. Loeb **1996**. Tolerance of different proteins for amino acid diversity.*Mol. Diversity 2*: 111–118.
12. Tao, H. and V.W. Cornish **2002**. Milestones in directed enzyme evolution. *Curr. Opin. Chem. Biol. 6*: 858–864.
13. Zhao, H. and F.H. Arnold **1997**. Optimization of DNA shuffling for high fidelity recombination. *NARes 25*: 1307–1308.
14. Zhao, H., L. Giver, Z. Shao, J.A. Affholter and F.H. Arnold **1998**. Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nature Biotechnol. 16*: 258–261.

5 FACS Screening of Combinatorial Peptide and Protein Libraries Displayed on the Surface of *Escherichia coli* Cells

Thorsten M. Adams, Hans-Ulrich Schmoldt, and Harald Kolmar

5.1 Introduction

Screening of functional proteins by directed evolution requires the construction of a gene library (genotype), the establishment of a linkage between a protein and its encoding gene, and the selection of proteins of the desired function (phenotype) from the library. One elegant way of achieving phenotype/genotype coupling is to anchor the protein under consideration directly on the surface of the producing entity, be it a eukaryotic or bacterial cell [1,2], a virion [3], or even a ternary complex of sequestered mRNA, ribosome, and protein [4]. Such a system was first described by George Smith in 1985 [3], who fused peptides and proteins to the pIII protein of the filamentous phage M13. As an alternative to phage display, microbial cell surface display of the protein of interest has been introduced, which allows one to overcome two major drawbacks of phage display: (I) the necessity of reinfecting bacteria with the phage population obtained after each selection round, bearing the risk of losing diversity and (II) the unfeasibility of using fluorescence-activated cell sorting (FACS) as a highly efficient tool to screen a population of variants for ligand binding, since virions are too small to be detected by the LASER optics. Various expression systems have been developed for the display of peptides and proteins on the surface of *E. coli*, which is the most suitable host for the creation, breeding, and maintenance of large molecular repertoires that may be derived from over 10^{10} individual transformants. For other organisms, larger efforts are necessary to achieve comparable library sizes, due to lower transformation efficiencies than in *E. coli*. Meanwhile, a potpourri of different *E. coli* display formats is available (for review see, e. g., [5,6]). No predictions, however, can yet be made as to which display system suits the particular needs best. Several parameters need to be considered, such as the folding kinetics of the passenger protein to be displayed, which can influence translocation through the cytoplasmic and the outer membrane, the desired number of passenger proteins residing on the surface of a single cell, or influences of high-level expression of the protein of interest on cell viability, which is especially important if large molecular repertoires are to be handled [7].

To be presented on the surface of a bacterial cell, the protein of choice, after having been synthesized in the cytoplasm, has to pass the cytoplasmic and the outer membrane. This is generally achieved by genetically fusing the passenger protein

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst
Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-30799-0

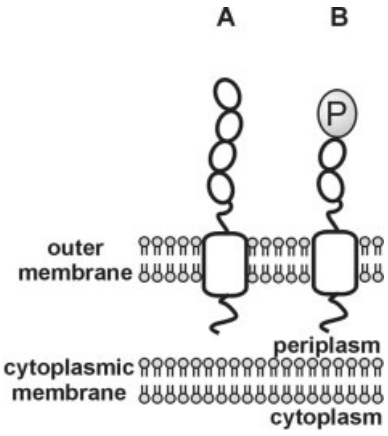


Fig. 5.1. (A) Schematic representation of EHEC intimin EaeA. (B) Schematic representation of truncated intimin (referred to as intimin′) lacking two carboxy-terminal domains and with a passenger domain fused to its carboxy terminus.

to a translocator domain that resides in the outer membrane and protrudes into the extracellular milieu. In this article, we focus on a surface display format based on a truncated version of intimin, an adhesin from pathogenic *E. coli*. The enterohemorrhagic *E. coli* (EHEC) intimin EaeA mediates adherence of the bacteria to eukaryotic target cells [8]. Intimin is composed of a transmembrane region of unknown structure, from which 3 immunoglobulin-like domains and 1 lectin-like domain protrude into the extracellular milieu (Fig. 5.1A). Those 4 extracellular domains form a rigid rod-like structure that is thought to be anchored to the transmembrane domain via a flexible hinge formed by 2 glycine residues [9]. Expression vector pASKInt100 contains the coding sequence for intimin′, a truncated version of intimin lacking the lectin-like domain and the C-terminal immunoglobulin-like domain (Fig. 5.1B). Intimin′ can be expressed in *E. coli* K12 and serves as a translocator and membrane anchor of the protein of interest fused to its carboxy terminus, which eventually becomes exposed on the cell surface remote from the lipopolysaccharide layer. With this system, we were able to display a variety of passenger domains, among them the protease inhibitor EETI-II, the Bence-Jones protein REI_v, interleukin-4 [7], ubiquitin, TEM-1 β-lactamase variants, β-lactamase inhibitor protein (BLIP), calmodulin, and peptides 50–70 amino acids long. Approximately 30 000 passenger proteins were found to be located on the surface of a single bacterium [7]. The overexpression of intimin′ fusion proteins is very well tolerated by the producing cell, which makes this system ideally suited for the handling of large libraries.

As indicated above, one major advantage of bacterial and other cell-based surface display formats lies in the ability to use fluorescence-activated cell sorting for high-throughput library screening. With modern FACS equipment, such as the Cytomation MoFlo or the FACSVantage from Beckton-Dickinson, sorting rates of up to 100 000 events per second are possible [10].

In many applications of bacterial surface display, the screening process is aimed at isolating a peptide or protein with an enhanced binding capability to a given ligand. To this end, cells displaying multiple copies of a particular peptide or protein variant on their surface are incubated with a fluorescently labeled ligand. Then,

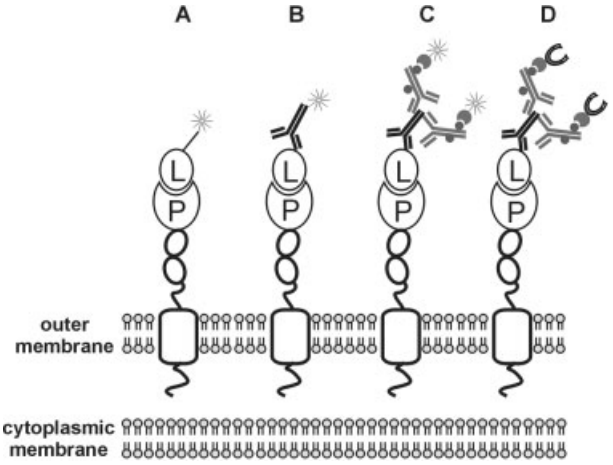


Fig. 5.2. Methods for detection of passenger-ligand interaction. (A) Fluorophore-coupled ligand; (B) fluorophore-coupled antibody; (C) quaternary complex generated by subsequent rounds of incubation with ligand, primary antibody, biotinylated second antibody, and streptavidin, R-phycoerythrin conjugate; (D) quaternary complex generated by subsequent rounds of incubation with ligand, primary antibody, biotinylated second antibody, and streptavidin-coated magnetobeads. L: ligand; P: passenger.

cells are washed thoroughly and subjected to FACS. Cells that fall within a positive window of fluorescence are isolated and propagated. This procedure is repeated for several rounds until clones with the desired binding properties are enriched. Several methods of fluorescence labeling are available (Fig. 5.2). Besides direct chemical coupling of a fluorophore, e. g., fluorescein isothiocyanate (FITC), to the ligand, the use of a fluorescently labeled antibody directed against the ligand is likewise possible. Labeling can also be done by consecutive rounds of incubation with an antibody directed against the ligand, a second biotinylated antibody and streptavidin, and R-phycoerythrin conjugate. The latter procedure bears the advantage of dramatic signal amplification, since one molecule of phycoerythrin contains 35 or more fluorophores [11]. Furthermore, protein ligands produced by heterologous gene expression can be tagged with an epitope sequence for which a monoclonal antibody is commercially available and can be used for indirect immunofluorescence staining.

For practical reasons, especially when handling libraries exceeding 10^8–10^9 different variants, it is advisable to use magnetic cell sorting (MACS) as a less time-consuming procedure for the enrichment of candidate clones. To achieve this, a small molecule like biotin or fluorescein is coupled to the target protein. After incubation of the cell population with the labeled target protein, unbound target is removed by washing and centrifugation, and the cells are incubated with paramagnetic microbeads that are coated with biotin-binding streptavidin or a fluorescein-binding antibody. A separating column is placed in a strong high-gradient magnetic field generated by a permanent magnet. Then the cell population is passed through

the column, and the labeled cells are retained in the column while the unlabelled cells are washed away [12]. For elution, the column is simply removed from the magnet and the cells are washed out. A single-pass enrichment ratio of over 1000-fold has been reported [13]. Over 10^{10} bacterial cells can be handled in parallel in a single experiment, thus allowing one to screen large repertoires with reasonable library oversampling (unpublished results).

Before getting started in generating large libraries of the protein to be displayed on the *E. coli* cell surface, one has to verify that the protein of interest is amenable to intimin′-mediated surface display. To investigate this, the corresponding gene has to be cloned into the display vector pASKInt100 (Fig. 5.3) using unique *Ava* I (*Sma* I, *Xma* I) and *Bam* HI restriction sites to obtain an in-frame gene fusion consisting of the intimin′ coding sequence and the gene of interest. Expression of the resulting bipartite gene fusion is under *tetA* P/O control in pASKInt100 and can be induced by anhydrotetracycline. The *intimin′* gene contains an *amber* stop codon at position 35 of the *eaeA* gene [7]. By using an *amber* suppressor strain like DH5α or 71-18, translational read-through occurs at a reduced frequency that allows net accumulation of the fusion at approximately 30 000 copies per cell, which normally does not negatively interfere with cell viability [5]. If high-level expression of the particular passenger domain results in a reduced survival rate, the total number of surface-exposed passenger molecules can be adjusted to a tolerable level by using a nonsuppressor strain as expression host which contains a helper plasmid encoding

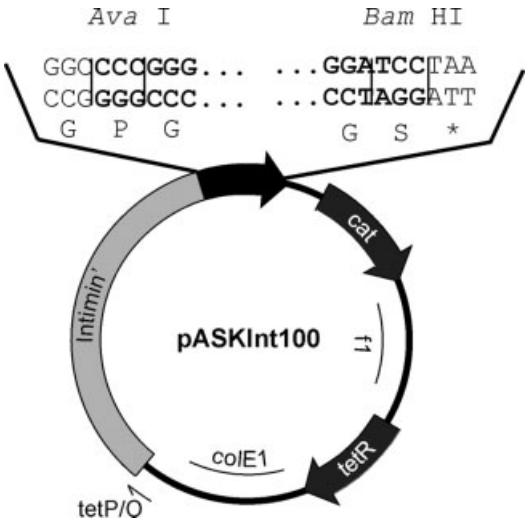


Fig. 5.3. Schematic representation of the display vector pASKInt100. f1, f1 replication origin; cat, chloramphenicol resistance marker; tetR, tetracycline repressor encoding gene; tetP/O, tetracycline promotor/operator region; colE1, ColE1 replication origin; intimin′, truncated *eaeA* gene of EHEC O157:H7. Unique *Ava* I (*Sma* I, *Xma* I) and *Bam* HI restriction sites allow the in-frame fusion of genes encoding various passenger domains, as described in further detail in Wentzel et al. [7].

the *supE* tRNA gene under *lac* promoter control. The concentration of inducer IPTG controls the number of suppressor tRNA molecules, which directly influences the efficiency of translational read-through at the *amber* codon [7].

Successful cell surface display of the protein can be verified by inducing gene expression via addition of anhydrotetracycline to the culture medium and immunofluorescence staining of the cells using an antibody directed against the protein to be displayed (Protocol 1). Analysis can be performed by fluorescence microscopy or flow cytometry.

In the following sections, we give detailed guidelines on the generation of peptide/protein libraries via construction of variants fused to intimin′, *E. coli* cell surface display, and library screening using magnetic (MACS) and fluorescence activated cell sorting (FACS).

5.2 Materials

5.2.1 *Escherichia coli* Strains and Plasmids

- 71-18: [F′ *lac*I^q(*lac*ZΔM15), *pro*A^+B^+; Δ*(lac-pro*AB*) sup*E*, thi*] (Source: B. Müller-Hill [14]).
- DH5α: [*endA1 hsdR17* ($r_{k^-}m_{k^-}$), *supE44*, *thi1*, *recA1*, *gyrA(Nal^{r}r)*, *relA1*, Δ*(lacZYA-argF)*U169, Φ*80lacZΔM15*] [15].
- Plasmid pASKInt100: Construction of pASKInt100 has been described in detail in Wentzel et al. [7]. The complete annotated nucleotide sequence of this vector is available at our website (http://www.gwdg.de/∼hkolmar).

5.2.2 Liquid Media and Agar Plates

- Chloramphenicol stock solution: 25 mg mL^{-1} in 96% (v/v) ethanol.
- 2YT medium: 1% (w/v) yeast extract; 1.6% (w/v) bacto tryptone; 0.5% (w/v) NaCl.
- 2YT-Cm^{25} medium: 2YT medium supplemented with 1/1000 vol of chloramphenicol stock solution added after autoclaving.
- 2YT-Cm^{25} plates: 2YT-Cm^{25} medium supplemented with 1.5% (w/v) bacto agar.
- M9 minimal medium plates: 0.7% (w/v) $Na_2HPO_4{\cdot}2H_2O$, 0.3% (w/v) KH_2PO_4, 0.1% (w/v) NH_4Cl, 1.5% bacto agar. After autoclaving, add the following sterilized solutions: 25 mL 20% (w/v) glucose, 1 mL 100 mM $CaCl_2$, 1 mL 1 M $MgSO_4$, 5 mL 0.1 mM $FeCl_3$, 1 mL thiamine solution (1 mg mL^{-1}).
- SOB medium: 2% (w/v) bacto tryptone, 0.5% (w/v) yeast extract, 0.05% (w/v) NaCl.
- SOC medium: 2% (w/v) bacto tryptone, 0.5% (w/v) yeast extract, 10 mM NaCl, 2.5 mM KCl. After autoclaving add the following sterilized solutions at the final concentration indicated: 10 mM $MgCl_2$, 10 mM $MgSO_4$, 20 mM glucose.

5.2.3 Biological and Chemical Materials

- 7 M ammonium acetate.
- 96% (v/v) ethanol.
- 10× T4 DNA Ligation Buffer (MBI Fermentas).
- Alexa Fluor® 488 (Molecular Probes, A-20000).
- Anhydrotetracycline stock solution: 2 mg mL^{-1} anhydrotetracycline (Acros) in N,N-dimethylformamide.
- Anti-mouse IgG (whole molecule)–biotin conjugate (Sigma-Aldrich, B-7264).
- Anti-rabbit IgG (whole molecule)–biotin conjugate (Sigma-Aldrich, B-7389).
- Bovine serum albumin, BSA 100×, 10 mg mL^{-1} (New England Biolabs).
- Dimethylsulfoxide (Sigma-Aldrich).
- EZ-Link Sulfo-NHS-Biotin (Pierce Biotechnology, 20217).
- HABA/Avidin reagent (Sigma-Aldrich, H-2153).
- N,N-dimethylformamide (Fluka).
- Paraffin.
- Phenol/chloroform 1:1.
- PBS: 140 mM NaCl, 10 mM KCl, 6.4 mM $Na_2HPO_4 \cdot 2H_2O$, 2 mM KH_2PO_4.
- Streptavidin-coated super-paramagnetic microbeads (Miltenyi Biotech).
- Streptavidin, R-phycoerythrin conjugate, 1 mg mL^{-1} (Molecular Probes).
- Sucrose gradient buffer: 100 mM Tris-HCl, pH 8.0, 10 mM EDTA, 100 mM NaCl.
- T4 DNA Ligase HC (high concentration), 30 U μL^{-1} (MBI Fermentas).

5.2.4 Equipment

- Fluorescence microscope Axioskop (Zeiss).
- Gene Pulser® (BioRad).
- Gradient mixer, 15 mL (MBI Fermentas).
- MidiMACS columns – stand and separation unit (Miltenyi Biotech).
- MoFlo FACS (Cytomation, Inc. or similar).
- Ultracentrifuge tubes 13.2 mL thin wall (Herolab).

5.3 Protocols

5.3.1 Verification of Cell Surface Exposure of the Passenger Protein

Protocol 1: Immunofluorescence Staining of *E. coli* Cells

1. Inoculate an Erlenmeyer flask containing 50 mL 2YT-Cm^{25} medium with 100 µL of a fresh overnight culture of *E. coli* strain DH5α or 71-18 carrying the pASK-Int100 plasmid that contains the gene coding for the protein to be displayed. Shake the culture flask at 37 °C.

2. At an OD_{600} of 0.2, add 5 µL of anhydrotetracycline stock solution to the culture medium. Shake at 37 °C for another 60 min.
3. Pellet cells (200–500 µL) by centrifugation in a tabletop centrifuge for 2 min.
4. Remove supernatant, resuspend the cell pellet in 10 µL PBS and add 1 µL (1 mg mL^{-1}) of an antibody directed against the protein to be displayed. Leave on ice for 10 min.
5. Wash the cells by addition of 180 µL PBS and centrifugation as in step 3.
6. Remove the supernatant and resuspend the cells in 10 µL PBS containing anti-mouse or anti-rabbit IgG biotin conjugate as required (1:10 dilution). Leave on ice for 10 min and wash the cells by addition of 180 µL PBS and centrifugation.
7. Resuspend the cell pellet in 10 µL of PBS containing streptavidin, R-phycoerythrin conjugate (1:10 dilution) and incubate on ice again for 10 min.
8. Wash the cells by addition of 180 µL PBS and centrifugation and resuspend the pellet in 10 µL PBS.
9. Analyze by fluorescence microscopy (Zeiss filter number 15, absorption: 565 nm, emission: 578 nm) or by flow cytometry (Fl2).

5.3.2 Labeling of the Target Protein

The target molecule can be directly labeled with a fluorescent dye or it can be indirectly labeled via biotinylation and successive incubation with a streptavidin, R-phycoerythrin conjugate. We recommend to biotinylate the target protein since it can be used both for MACS and for FACS then. For direct fluorescence labeling of the target molecule, NHS-coupled fluorescence dyes, e. g., Alexa Fluor® 488 from Molecular Probes, can be used, following the procedure described in protocol 2.

Protocol 2. Biotinylation of Target Protein using EZ-Link Sulfo-NHS-Biotin

1. Dissolve the target protein in PBS at approximately 1 mg mL^{-1}.
2. Add the corresponding amount of EZ-Link Sulfo-NHS-Biotin, to get a 10-fold molar excess over the target protein, directly to the protein solution. Incubate at room temperature for 30 min or on ice for 2 h.
3. Remove unreacted biotin molecules by dialysis against PBS.
4. Determine the biotin/protein ratio, which should be in the range of 3:1 to 5:1, using HABA-avidin reagent according to the instructions at the Pierce website (http://www.piercenet.com). If the coupling efficiency was not sufficient, try a larger molar excess of NHS-biotin over the target protein, up to 100-fold.

5.3.3 Library Construction

5.3.3.1 Vector Preparation

Protocol 3. Preparation of Vector DNA

To isolate pASKInt100 plasmid DNA, any convenient midi-prep purification system (e. g., Qiagen® Plasmid Midi Kit) can be used. It is advisable to extract the purified

DNA once with phenol/chloroform and once with chloroform, and to precipitate the DNA by addition of 1/10 volume of 7 M ammonium acetate and 3 volumes of 96% ethanol. Dissolve the DNA in H_2O at approximately 1 $\mu g\,\mu L^{-1}$.

1. Digest 50–100 µg pASKInt100 vector DNA in a total volume of 200 µL with 2–3 units of *Ava I* per µg DNA at 37 °C for 5 h or overnight in the manufacturer's recommended buffer.
2. After monitoring *Ava I* cleavage by agarose gel electrophoresis, extract digested DNA once with phenol/chloroform and once with chloroform, and precipitate it by adding 1/10 vol of 7 M ammonium acetate and 3 volumes of 96% (v/v) ethanol. Incubate for at least 30 min at −20 °C.
3. Centrifuge in a tabletop centrifuge for 15–30 min. Remove the supernatant completely, dry the pellet at 37 °C, and resuspend it in an appropriate volume of H_2O for *Bam* HI digestion.
4. Digest the DNA in a final volume of 200 µL with 2–3 units of *Bam* HI per µg DNA at 37 °C for 5 h or overnight in the manufacturer's recommended buffer.
5. Prepare solutions of 40%, 30%, and 10% sucrose (w/v) in sucrose gradient buffer. Place 500 µL of 40% sucrose solution in a 13.2-mL ultracentrifuge tube. Layer a 10%–30% sucrose gradient on top of the 40% cushion, using a gradient mixer.
6. Layer the DNA solution from the restriction digest on top of the sucrose gradient. Fill the tube with paraffin to approximately 3 mm below the rim. Balance with paraffin.
7. Centrifuge at 30 000 rpm for 21 h in a type TST41.14 rotor (Kontron, swing-out) or equivalent at 15 °C.
8. Fractionate the gradient in 500 µL aliquots by puncturing the bottom of the tube with a needle and collecting drops after removing the needle.
9. Analyze the fractions by agarose gel electrophoresis. Combine the fractions containing the vector fragment and precipitate the DNA with 1/10 volume of 7 M ammonium acetate and 3 volumes of 96% (v/v) ethanol.

5.3.3.2 Insert Generation

The techniques for generating repertoires of a given protein to be displayed are well established. The particular technique to be used depends on the conceptual formulation of the experiment. If a certain region of the passenger protein, e. g., a binding loop, is planned to be mutated, with the aim of changing or enhancing the binding abilities of the protein, the corresponding wild-type sequence can be replaced by a synthetic cassette of degenerated oligonucleotides.

Alternatively, the whole gene can be mutated by DNA shuffling [16] or error-prone PCR [17]. In any case, the gene library should be designed so that it contains DNA ends compatible with *Ava* I and *Bam* HI restricted pASKInt100.

5.3.3.3 Ligation and Transformation

Ligation of digested vector and insert DNA is performed by standard procedures. For creating large libraries it is advisable to set up several ligation reactions in parallel.

Protocol 4. Ligation

1. Set up 12 ligation reactions, each containing 300 ng of digested vector DNA, 3-fold molar excess of insert DNA, 2 µL 10× T4 DNA ligase buffer, 0.2 µL BSA (100×), 1 µL T4 DNA Ligase (HC), and H_2O to 20 µL.
2. Incubate overnight at 15 °C.
3. Inactivate ligase by incubation at 65 °C for 10 min.
4. Pool the 12 reactions and extract once with 1 volume of phenol/chloroform and once with 1 volume of chloroform. Precipitate the DNA with 1/10 vol of 7 M ammonium acetate and 3 vol of 96% (v/v) ethanol. Incubate for at least 30 min at −20 °C, then centrifuge in a tabletop centrifuge for 15–30 min.
5. Discard the supernatant and resuspend the DNA pellet in 240 µL H_2O and either store at −20 °C or use directly for electroporation.

Transformation is done by electroporation of 71-18 or DH5α cells. According to our experience, the use of chemically competent cells is not recommended, because it gives much lower yields of transformants.

Protocol 5. Electroporation of *E. coli* Cells

The following materials have to be prepared in advance: 2 L of sterile double-distilled H_2O pre-cooled on ice; 36 SOC-Cm^{25} agar plates in large petri dishes (15 cm in diameter); 10 SOC-Cm^{25} agar plates in small petri dishes (9.2 cm in diameter); two 1-L Erlenmeyer flasks each containing 400 mL SOB medium, one flask containing 100 mL SOC medium, and one flask containing 100 mL SOB medium.

A. Preparation of electrocompetent cells

1. Inoculate 5 mL SOB medium with DH5α or 71-18 cells grown on a M9 minimal-medium plate and shake overnight at 37 °C.
2. Inoculate each of the two flasks containing 400 mL SOB medium with 2 mL of the overnight culture. Shake at 37 °C until an OD_{600} of 0.5 is reached (∼2–3 h).
3. Partition the cell culture into 16 50-mL plastic tubes (Falcon) and centrifuge at 4000 rpm for 10 min at 4 °C (Hettich Rotixa/RP or equivalent). Discard the supernatant.
4. Carefully resuspend the cell pellets each in 50 mL of pre-cooled H_2O (on ice). Incubate on ice for 1 h.
5. Centrifuge at 4000 rpm for 10 min at 4 °C. Discard the supernatant.
6. Wash as in step 4 with 25 mL of pre-cooled H_2O. Incubate on ice for 1 h and centrifuge as in step 5. Discard the supernatant.

7. Wash as in step 4 with 10 mL of pre-cooled H_2O. Distribute the contents of 4 plastic tubes equally into the remaining 12 tubes. Leave on ice for 1 h and centrifuge as in step 5. Discard the supernatant. Place the tubes on ice for another 10 min and resuspend the cells in the remaining liquid, which should not exceed 200 µL.

B. Electroporation

1. Add 20 µL of the ligation reaction to each of the 12 aliquots of competent cells and incubate for at least 30 min on ice, then transfer into pre-chilled electroporation cuvettes.
2. Set up a Gene Pulser to give a 2500 V pulse, using a 25 µF capacitor, and adjust the resistance to 200 Ω. Place the cuvette into the electroporation chamber and pulse once.
3. Immediately add 800 µL SOC medium and transfer the cell suspension into a test tube. Rinse the cuvette twice with 800 µL SOC medium and incubate the cells with agitation at 37 °C for 45–60 min.
4. Pool the contents of all 12 tubes. Remove an aliquot and make a serial dilution to determine the total number of transformants (10^6–10^9 can be expected when setting up 12 ligation reactions). Plate the dilutions on the small SOC-Cm^{25} agar plates. Streak the remaining cell suspension on the 36 large SOC-Cm^{25} agar plates and incubate overnight at 37 °C.
5. Count the colonies on the dilution plates the next day to determine the total number of transformants. Harvest the library cells by flooding each of the 36 plates with 4 mL SOC medium and detach cells by scraping off under sterile conditions. Pool the cell suspension and add DMSO to a final concentration of 9% (v/v). Store at −70 °C in, e. g., 2-mL aliquots or use directly in library screening.

5.3.4 Combinatorial Library Screening by FACS and MACS

This section describes the isolation of a protein variant with desired binding capabilities to a given ligand protein, starting with a large repertoire exceeding 10^9 primary transformants. It is advisable to screen at least 10 times more cells than primary transformants that have been obtained, which would take more than 24 h in the first round using FACS even with high-speed cell sorters. Therefore it is recommended to pre-enrich clones that interact with the ligand by using MACS prior to FACS.

Protocol 6. Pre-enrichment by MACS

1. Inoculate an Erlenmeyer flask containing 500 mL 2YT-Cm^{25} with an aliquot of the thawed DMSO culture (∼100 µL) representing the library clones. Shake at 37 °C until an OD_{600} of 0.2 is reached.

2. Add 50 µL of anhydrotetracycline stock solution. Shake the culture flask at 37 °C for another 60 min. Place on ice for 4 h.
3. Centrifuge at 4000 rpm for 30 min (Hettich Roto Silenta/RP or equivalent) at 4 °C. Discard the supernatant. Resuspend the cell pellet in 40 mL PBS, transfer the cell suspension to a 50-mL tube, and centrifuge at 4000 rpm for 10 min at 4 °C (Hettich Rotixa/RP or equivalent). Discard the supernatant.
4. Resuspend the cell pellet in 2.5 mL PBS and add the ligand protein to a final concentration of 1 µM (or lower to get higher-affinity binders). Incubate on ice for 20 min. Gently shake every 5 min.
5. Add PBS to 50 mL and centrifuge at 4000 rpm for 10 min at 4 °C (Hettich Rotixa/RP or equivalent). Discard the supernatant.
6. Resuspend the cell pellet in 2.5 mL PBS and add 40 µL streptavidin-coated super-paramagnetic microbeads. Incubate for 15 min on ice. Mix gently every 5 min. Wash as in step 5.
7. Resuspend the cells in 40 mL PBS. Remove an aliquot to make a serial dilution to determine the total number of living cells by plating onto small 2YT-Cm^{25} agar plates (should be approximately 10^{10} cells).
8. Apply 5 mL of the cell suspension to a MidiMACS column equilibrated with 5 mL PBS, which is placed in a permanent magnet (8 columns are required all together).
9. Wash 3 times with 5 mL PBS.
10. Remove the column from the magnetic field and elute with 5 mL PBS. Pool the eluted cells from all 8 MACS columns and centrifuge at 4000 rpm for 10 min at 4 °C (Hettich Rotixa/RP or equivalent). Discard the supernatant, resuspend the cells in 5 mL PBS and apply them to another MidiMACS column. Wash as in step 9 and elute with 3 mL PBS. Remove an aliquot of the eluted cell suspension to make a serial dilution to determine the total number of cells that have been retained on the column and plate on small 2YT-Cm^{25} agar plates. Plate the remaining suspension on 3 large 2YT-Cm^{25} agar plates. Incubate overnight at 37 °C.
11. Count the colonies on the dilution plates to determine the enrichment factor, which is the ratio of the input and output cell numbers. Harvest the colonies on the large agar plates by flooding each plate with 4 mL SOC medium and detach cells by scraping off under sterile conditions. Add DMSO to a final concentration of 9% (v/v). Store at −70 °C or use directly for the next screening round.

For the second round of MACS enrichment, the above protocol can be scaled down to 50 mL of induced culture and one MidiMACS column.

Protocol 7. FACS

1. Inoculate a flask containing 50 mL 2YT-Cm^{25} medium with an aliquot (∼50 µL) of the DMSO culture of the pre-enriched library. Shake the culture flask at 37 °C.

2. At an OD_{600} of 0.2, add 5 µL of anhydrotetracycline stock solution. Shake the culture flask at 37 °C for another 60 min.
3. Pellet cells (200–500 µL) by centrifugation in a tabletop centrifuge for 2 min.
4. Resuspend the cell pellet in 10 µL of PBS and add the ligand protein at a final concentration of 1 µM. For selection of higher-affinity binders, the concentration may be further reduced. Incubate on ice for 20 min.
5. Wash the cells by adding 180 µL PBS and centrifuge in a tabletop centrifuge for 2 min.
6. If the ligand protein is fluorescently labeled, resuspend the cells in 10 µL PBS and subject directly to FACS. If the ligand protein is biotinylated (protocol 2), resuspend the cells in 10 µL PBS containing streptavidin, R-phycoerythrin conjugate (1:10 dilution in PBS). Incubate on ice again for 10 min.
7. Wash the cells by adding 180 µL PBS and centrifuge as in 5.
8. Resuspend the cells in 10 µL PBS and subject to FACS.
9. Sort cells with a Cytomation MoFlo cell sorter with the following parameters: forward scatter, side scatter, 730 (LIN mode, amplification factor 6); FL1, 600 (LOG mode); FL2, 600 (LOG mode); trigger parameter, side scatter. The sample flow rate should be adjusted to an event rate of approximately 30 000 s^{-1}. Adjust the sorting gate so that approximately 0.1% of the cells fall within the positive window.
10. Remove an aliquot, make a serial dilution and plate on small 2YT-Cm^{25} agar plates to determine the survival rate. Plate the sorted cells on 1 large 2YT-Cm^{25} agar plate. Incubate overnight at 37 °C.
11. Harvest the colonies by flooding the plate with 2YT or SOC medium and detach cells by scraping off under sterile conditions. Add DMSO to a final concentration of 9% (v/v). Store at −70 °C or use directly for next screening round.

After 3–4 rounds of combined MACS/FACS screening, enrichment of positive clones should be detectable. If a significant portion of the population is FACS-positive, individual clones can be labeled and analyzed as described in Protocol 7. A control should be included, in which the labeling procedure is performed without the ligand protein, to ensure that the observed interaction is ligand-specific.

5.4 Troubleshooting

After construction of the intimin′ fusion protein, the successful cell surface exposure of the passenger protein should be verified. This can be achieved either by extending the passenger protein by an epitope sequence for which an antibody is commercially available or, as outlined in Protocol 1, by labeling the induced cells with an antibody against the passenger protein. If no surface exposure can be detected, translocation of the passenger through the outer membrane may be ham-

pered. Most likely, the passenger protein has to traverse the outer membrane in an unfolded state (Adams et al., submitted for publication). If periplasmatic folding of the passenger occurs too quickly, membrane translocation can be hindered. In this case, it may be helpful to add agents to the growth medium that could influence protein folding. If the passenger protein contains disulfide bonds, addition of 20 mM β-mercaptoethanol can help achieve cell-surface exposure. Using an *E. coli* strain lacking the major periplasmic oxidoreductase DsbA, which promotes periplasmic disulfide bond formation, can also improve the display of the passenger. Adding sodium dodecyl sulfate (SDS) to a final concentration of 0.6% to the growth medium was found to be helpful in the display of some proteins (Adams et al., submitted for publication). SDS should be washed away before induction of gene expression to avoid SDS binding to the intimin′-passenger protein fusion.

In principle, any cell sorter can be used for library screening, although it is advantageous to have a high-speed cell sorter available to be able to screen large repertoires with reasonable oversampling. With modern devices, approximately 10^8 clones can be screened in one hour, which allows one to sort an ensemble of 10^7 individual variants with 10-fold oversampling. For larger repertoires, MACS should be used to decrease the number of unwanted cells by a factor of 100–1000 in a single sorting round. If a high-speed sorter is not available, using MACS is mandatory when dealing with large repertoires. However, in contrast to FACS, the enrichment of cells cannot be directly monitored. Therefore, it is advisable to count the cells before and after MACS by plating serial dilutions to determine the enrichment ratio.

One problem that may arise is the selection of clones that bind to streptavidin, R-phycoerythrin conjugate. This conjugate is abundantly used in the labeling schemes described above. Therefore, enrichment of false-positive clones can easily occur. The solution to this problem is to use an alternating labeling scheme. For example, the first screening round can be carried out with a biotinylated ligand, and ligand binding is detected with streptavidin, R-phycoerythrin conjugate; the next sorting round can be carried out with a FITC-coupled ligand or an FITC-coupled antibody directed at the ligand.

Candidate binders that have been isolated by MACS/FACS can be conveniently analyzed further by FACS. Procedures for measuring the dissociation rate constants and the association rate constants have been described [18,19]. If the selection yields low-affinity binders only, the stringency can be further increased by reducing the concentration of the ligand, by extending the time of incubation after labeling to select for binders with a lower dissociation rate constant, or by reducing the incubation time for cell-ligand interaction to select binders with a higher association rate constant.

5.5 Major Applications

An important application is the isolation of proteins with a higher affinity to a given ligand. Georgiou and coworkers displayed variants of a single-chain antibody (scF_v) on the *E. coli* cell surface and screened for clones binding to a fluorescently labeled hapten by FACS [20]. The authors were able to isolate an affinity-improved clone having a K_D about threefold lower than the wild-type in a single sorting round. Bacteria displaying products of a certain gene or genome can be used for epitope mapping and the isolation of monospecific antibodies. This is exemplified by the work of Christmann et al., who displayed random fragments derived from the classical swine fever virus (CSFV) envelope protein E(rns) on the surface of *E. coli* [21]. By incubation of cells with polyclonal anti-E(rns) serum, a major linear antigenic determinant of the E(rns) protein could be identified and linear epitopes could be mapped. Moreover, strategies have been developed to use bacterial surface display and FACS for enzyme engineering using fluorescent enzyme substrates [22].

Besides its application for high-throughput screening of combinatorial peptide, protein, and enzyme libraries, bacterial surface display of heterologous proteins has proven to be a useful strategy for several other applications. For example, *E. coli* cells have been constructed to serve as bioadsorbents for heavy metals. To this end, 2 hexahistidine clusters that were fused to a permissive loop of LamB increased Cd^{2+} sequestration by recombinant *E. coli* 11-fold and enabled the cells to adhere reversibly to a Ni^{2+}-containing solid matrix in a metal-dependent manner [23]. By fusing a mouse metallothionein to *N. gonorrhoeae* autotransporter and displaying this construct on the surface of metal-tolerant *Ralstonia eutropha*, Valls et al. were able to increase the Cd^{2+} binding ability of this strain three-fold [24]. Another application is the use of cells displaying a heterologous fusion protein on their surface directly as vaccines. Liljeqvist and colleagues engineered nonpathogenic staphylococci to display a functional cholera toxin B subunit (CTB) from *Vibrio cholerae* and proposed their use as live bacterial vaccines [25].

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft through SFB416 and the Mukoviszidose e.V.

References

1. K.D. Wittrup, *Curr. Opin. Biotechnol.* **2001**, *12*, 395–399.
2. G. Georgiou, C. Stathopoulos, P.S. Daugherty, A.R. Nayak, B.L. Iverson and R. Curtiss, *Nature Biotechnol.* **1997**, *15*, 29–34.

3. G.P. Smith, *Science* **1985**, *228*, 1315–1317.
4. C. Schaffitzel, J. Hanes, L. Jermutus and A. Pluckthun, *J. Immunol. Methods* **1999**, *231*, 119–135.
5. A. Hayhurst and G. Georgiou, *Curr. Opin. Chem. Biol.* **2001**, *5*, 683–689.
6. S.Y. Lee, J.H. Choi and Z. Xu, *Trends Biotechnol.* **2003**, *21*, 45–52.
7. A. Wentzel, A. Christmann, T. Adams and H. Kolmar, *J. Bacteriol.* **2001**, *183*, 7273–7284.
8. M. Batchelor, S. Prasannan, S. Daniell, S. Reece, I. Connerton, G. Bloomberg, G. Dougan, G. Frankel and S. Matthews, *EMBO J.* **2000**, *19*, 2452–2464.
9. Y. Luo, E.A. Frey, R.A. Pfuetzner, A.L. Creagh, D.G. Knoechel, C.A. Haynes, B.B. Finlay and N.C. Strynadka, *Nature* **2000**, *405*, 1073–1077.
10. R.G. Ashcroft and P.A. Lopez, *J. Immunol. Methods* **2000**, *243*, 13–24.
11. S. Ritter, R.G. Hiller, P.M. Wrench, W. Welte and K. Diederichs, *J. Struct. Biol.* **1999**, *126*, 86–97.
12. A. Christmann, K. Walter, A. Wentzel, R. Kratzner and H. Kolmar, *Protein Eng.* **1999**, *12*, 797–806.
13. Y.A. Yeung and K.D. Wittrup, *Biotechnol. Prog.* **2002**, *18*, 212–220.
14. U. Ruther and B. Muller-Hill, *EMBO J.* **1983**, *2*, 1791–1794.
15. R. Lutz and H. Bujard, *Nucleic Acids Res.* **1997**, *25*, 1203–1210.
16. W.P. Stemmer, *Nature* **1994**, *370*, 389–391.
17. R.C. Cadwell and G.F. Joyce, *PCR Methods Appl.* **1992**, *2*, 28–33.
18. L.A. Sklar, D.A. Finney, Z.G. Oades, A.J. Jesaitis, R.G. Painter and C.G. Cochrane, *J. Biol. Chem.* **1984**, *259*, 5661–5669.
19. E.T. Boder, K.S. Midelfort and K.D. Wittrup, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 10701–10705.
20. P.S. Daugherty, G. Chen, M.J. Olsen, B.L. Iverson and G. Georgiou, *Protein Eng.* **1998**, *11*, 825–832.
21. A. Christmann, A. Wentzel, C. Meyer, G. Meyers and H. Kolmar, *J. Immunol. Method* **2001**, *257*, 163–173.
22. M.J. Olsen, D. Stephens, D. Griffiths, P. Daugherty, G. Georgiou and B.L. Iverson, *Nature Biotechnol.* **2000**, *18*, 1071–1074.
23. C. Sousa, A. Cebolla and V. de Lorenzo, *Nature Biotechnol.* **1996**, *14*, 1017–1020.
24. M. Valls, S. Atrian, V. de Lorenzo and L.A. Fernandez, *Nature Biotechnol.* **2000**, *18*, 661–665.
25. S. Liljeqvist, P. Samuelson, M. Hansson, T.N. Nguyen, H. Binz and S. Stahl, *Appl. Environ. Microbiol.* **1997**, *63*, 2481–2488.

6 Selection of Phage-displayed Enzymes

Patrice Soumillion

6.1 Introduction

In 1985, G.P. Smith reported for the first time the expression of a foreign peptide at the surface of a filamentous bacteriophage [1]. Less than twenty years later, several thousands of publications, reporting mainly the discovery of peptides and antibodies that bind to a variety of receptors or ligands, attest to the tremendous success of phage display, which is now a well established, powerful technology [2]. The key of this success comes from the basic principle of physically linking an accessible expression product to its genetic information via a phage particle. This is achieved by simply cloning the foreign gene in fusion with a gene encoding a coat protein of the virion. Upon phage morphogenesis, the fusion protein is assembled in the phage, resulting in a chimeric particle displaying the foreign peptide or protein at its surface. Libraries of mutant phages are created by cloning libraries of genes that either come from natural sources or are constructed by random mutagenesis approaches. Then, phages displaying specific peptides or proteins that bind to a chosen target are easily selected by affinity capture, using an immobilized version of the target. The selected phages can be replicated and amplified by simple infection.

Although the selection of new binders (ligands) is the most straightforward application of phage display, the technology is also used for the directed evolution of enzymes [3]. Libraries of phage-displayed enzymes can be created, but the selection of mutants endowed with new catalytic properties requires more elaborated strategies, because it cannot be based solely on the binding of a substrate or an inhibitor. During the past decade, several methods have been developed with this aim. One of them is affinity capture using transition-state analogs (TSAs) as targets [4]. Indeed, because TSAs are supposed to mimic the geometry and charge distribution of transition states, a protein that binds to a TSA should also decrease the activation barrier of the corresponding reaction and therefore be a catalyst. A second indirect strategy takes advantage of 'suicide' substrates that covalently label the active sites of enzymes. These molecules are also called mechanism-based inhibitors, because they are transformed by the catalytic machinery of the enzyme into very reactive compounds that ultimately and irreversibly block the active site. In the selection protocol, the phage-enzymes are incubated under kinetic control with

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst
Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-30799-0

a limiting amount of a biotinylated suicide substrate. The most active enzymes are labeled faster and can subsequently be affinity-captured with immobilized streptavidin. This method was described initially for the selection of a phage displaying β-lactamase activity [5]. An improved protocol has since been developed to avoid selecting proteins that can react with the suicide substrate but do not turn over the substrate [6]. In this protocol, the library is incubated with a nonbiotinylated substrate to block all the active sites that are unable to turn over before labeling with the biotinylated suicide substrate. Other examples are the selection of glycosidase and phosphatase activities by using substrates immobilized through biotinylated suicide leaving groups [7, 8]. Phosphonylating agents – although not strictly suicide substrates – have also been used for selection of esterases and proteases [9, 10].

More recently, various methods using substrates for selection have been reported. In some of them, phages displaying metalloenzymes are immobilized via a substrate in the absence of the metal ion [11, 12]. Upon addition of the metal ion, the enzyme is activated and substrate turnover leads to elution of the phage. Other strategies use phages displaying both the enzyme and the substrate in such a manner that 'intraphage' turnover can take place [13–15]. Phages displaying the reaction product are then selected with an immobilized binder that specifically recognizes that product.

As illustrated in Figure 6.1, all these methods share many similarities from an experimental point of view. Hence, we will describe the selection strategy, using suicide substrates as an example, and complete the description with specific details regarding the other strategies. Usually, the objective is to evolve an existing enzyme towards a new or an improved catalytic property. The first step consists in cloning the gene for the enzyme into a phage-display vector and characterizing the properties of the phage-enzyme as well as possible. Then, a library of mutants is constructed by using a random mutagenesis method. Several methods, such as cloning degenerate oligonucleotides, error-prone PCR, and DNA shuffling are available, and the choice depends on the nature of the library that is wanted. Before starting the selections, the library should be characterized. Finally, the selections are carried out and, again, the choice of the method depends on the nature of the desired activity. However, in most instances, the method involves the use of a small synthetic molecule (substrate or inhibitor), often containing a labeling or capture module such as biotin. Selection protocols comprise several steps, not always in the same order, but often including incubation and/or reaction with the substrate or inhibitor in solution or immobilized, removal of the excess molecule, capture on a support, washings, and elution. Ideally, a model selection should be performed using a mixture of active and inactive phages, to determine the working conditions for the selection.

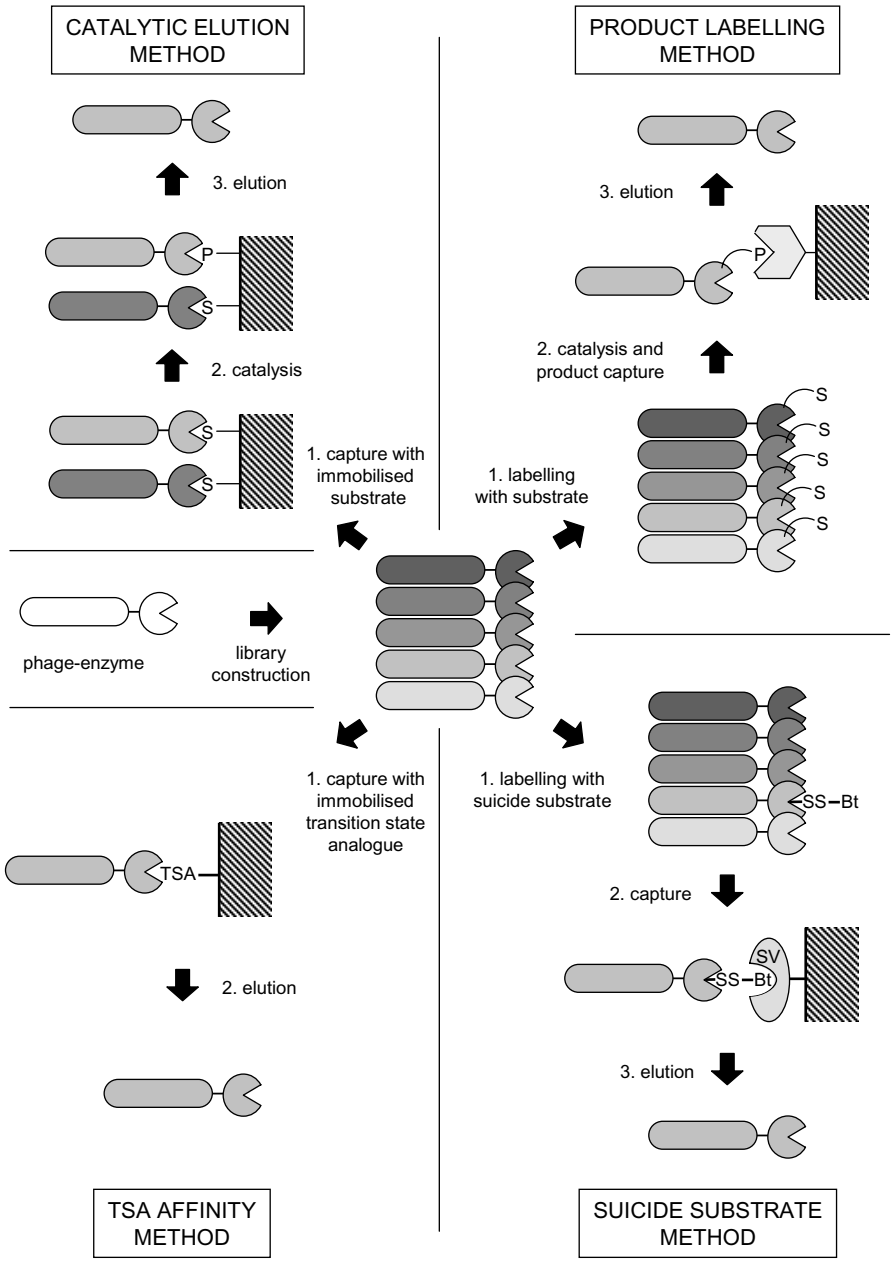


Fig. 6.1. Summary of methods for selecting phage-displayed enzymes on the basis of catalytic activity (S: substrate; P: product; SS: suicide substrate; Bt: biotin; SV: streptavidin; TSA: transition state analogue).

6.2 Materials

6.2.1 Buffers, Reagents and Consumables

- TBS: 50 mM Tris, 150 mM NaCl, pH 7.5
- MTBS: TBS containing 2% nonfat skim milk powder (BioRad, Nazareth Eke, Belgium)
- TTBS: TBS containing 0.1% Tween 20
- TE: 10 mM Tris, 1 mM EDTA, pH 8.0
- Phosphate 50 mM, pH 7.0
- 20% PEG 5000 (w/v)/2.5 M NaCl
- Dimethylsulfoxide (DMSO) for molecular biology (Fluka, Taufkirchen, Germany), dried on molecular sieve (5 Å)
- Oligonucleotides from Eurogentec (Liège, Belgium); degenerate oligonucleotides purified by polyacrylamide gel electrophoresis (PAGE)
- Restriction enzymes, T4 DNA ligase, and polymerases (e. g., from New England Biolabs, Beverly, MA, USA), as well as factor Xa, trypsin and bovine serum albumin (BSA, highest purity; e. g. from Sigma, Bornem, Belgium)
- Dynabeads M-270 streptavidin-coated magnetic beads from Dynal (Oslo, Norway)
- Anti-pIII antibody from MoBiTec (Göttingen, Germany)
- LB broth and LB agar from Invitrogen (Carlsbad, CA, USA)
- Large petri dishes ($30 \times 30\,cm^2$) from Nunc (Wiesbaden, Germany)
- 0.45 µm Millex-HV filters (Millipore, Billerica, MA, USA) and 1 µm Puradisc 25 GD prefilters (Whatman, Kent, UK)
- All other products and reagents from Sigma or Fluka
- Suicide substrates and other biotinylated molecules must be synthesized individually

6.2.2 Strains and Vectors

- Phage-enzymes were constructed by cloning genes into the vector fd-DOG1, a phage carrying the tetracycline resistance gene [16]. If expression control of the fusion protein is necessary, a phagemid vector such as pHDi.Ex should be used [17]. Many other vectors are also available [18–20].
- Two *E. coli* strains are used for phage infection and production. TG1 is a fast-growing strain, and JM 109 is used only if the modified phage shows genetic instability. JM109 carries the *recA* mutation, due to which the bacteria grow more slowly, and the phage yields are lower. The genotypes are:

  TG1: *supE thi-1* Δ*(lac-proAB)* Δ*(mcrB-hsdSM)5* (r_{K^-} m_{K^-}) [F′ *tra36 proAB* $lacI^q Z(M15]$

  JM109: e14$^-$(McrA$^-$) *recA1 endA1 gyrA96 thi-1 hsdR17*(r_{K^-} m_{K^-}) *supE44 relA1* Δ*(lac-proAB)* [F′ *tra36 proAB* $lacI^q Z\Delta M15$]

6.3 Protocols

Most of the protocols that are typically used in phage display can be adopted when working with phage-enzymes. A plethora of useful background information and current techniques is already available in several laboratory manuals dedicated to phage display [18–20]. These are excellent guides for anyone who is interested in the construction of phage-display libraries and the selection of specific binders. In this chapter, special emphasis is given to protocols, experimental aspects, and tricks related to the manipulation of phage-enzymes and to the selection steps involving catalytic reactivity.

6.3.1 The Phage-enzyme

6.3.1.1 Choosing a Vector

Several vectors have been developed for the display of proteins on filamentous phages such as M13 and fd. The coat of these phages is made of several thousand copies of a small but major protein (pVIII) and 3–5 copies of each of 4 minor proteins that are located at the tips of the filaments (pIII and pVI at one tip, pVII and pIX at the other). Although each of these proteins have been used for the display of foreign polypeptides, pIII is generally chosen for enzyme display. The pIII protein (406 amino acids, 42.5 kDa) is made of 3 domains, N1 (1–68), N2 (87–217), and CT (257–406) that are connected by glycine-rich sequences. N1 and N2 are necessary for phage infection, and CT, also called the anchoring domain, is essential for forming a stable phage particle.

The pIII protein is expressed as a precursor having an amino-terminal signal peptide necessary for addressing the protein through the periplasm of *E. coli*. The signal peptide is removed by a specific protease after secretion, and the pIII ends up anchored in the bacterial inner membrane. Its assembly in the phage particle is concomitant with phage extrusion.

Displaying an enzyme necessitates cloning its gene between the sequences encoding the signal peptide and the mature pIII. Two types of pIII-display vectors are commonly used, phages and phagemids, each with its own advantages and inconveniences. With phagemid vectors, the protein is expressed in fusion with either the full size or the carboxy-terminal (CT) domain of pIII. Phagemids have all the advantages of plasmids with regard to cloning, DNA manipulation, and control of expression with promoters. A stop codon could also be inserted between the two genes of the fusion. Therefore, expressing in a suppressive strain affords phage display, but the protein could also be expressed alone in nonsuppressive strains without the need of recloning. A disadvantage of phagemids is the relatively low level of surface display, which generally results in a large proportion of phage particles that do not display any fusion protein. This is due to competition between the fusion protein and the wild-type pIII encoded by the helper phage. This does

not occur with phage vectors for which the unique origin of the pIII protein is the fusion gene. Hence, higher surface display is generally obtained, with typical levels between 0.2 and 3 proteins per phage particle. Nevertheless, total display remains an exception, because *in vivo* proteolysis often partly removes the fusion protein from pIII. The absence of the helper phage infection step is also a practical advantage with phage vectors. They also contain antibiotic resistance markers, which allow for the detection of infected bacteria as colonies. Nevertheless, all the cloning and DNA manipulation must be performed using the replicative form (RF) of the phage, which is more difficult to produce in large amounts and high purity. Finally, the phage format does not allow the expression to be controlled at the DNA level. Although the choice is not obvious, we prefer phage vectors because, with phage-enzymes, a high level of display is usually a great advantage: it facilitates detection of activity and increases the sensitivity of selection. Sometimes, when low-affinity capture is necessary, multivalent display could be a prerequisite because of the need of avidity effects. This occurs, for example, when capturing enzymes on an immobilized substrate under conditions in which the enzyme is inactive [12].

The following protocols have been used for phage vectors, although most of them can be directly applied to phagemids.

6.3.1.2 Choosing a Linker

In phage display, the carboxy terminus of the protein of interest is fused to the amino terminus of the coat protein via a peptide linker. This linker must be of sufficient length for allowing proper folding of both proteins. Moreover, when the selection scheme involves capture by covalent trapping of the surface protein, a linker that is cleavable by a specific endoprotease is often interesting. Indeed, phage elution from the selecting support can then be easily performed by treatment with the specific protease. Finally, the linker should not be too susceptible to *in vivo* proteolysis, to prevent cleavage of the fusion protein during phage production.

Here are two examples of peptide linkers that we have successfully used with phage-enzymes:

- Ala-Ala-Ile-Glu-Gly-Arg-Ala-Ala: a linker that is cleavable by factor Xa endoprotease or trypsin and is quite resistant to *in vivo* proteolysis
- Gly-Gly-Gly-Ser-Gly-Gly-Gly-Ser: a noncleavable, flexible, hydrophilic linker that is highly resistant to *in vivo* proteolysis

6.3.1.3 Phage-enzyme Production

Phages are produced simply by growing infected bacteria. The quality and quantity of phages are not reproducible from one production to the next. Infected bacteria can produce phages at various temperatures, typically between 20 °C and 37 °C. The rate of phage production increases with increasing temperature and bacterial density. However, the level of surface display usually decreases with increasing growth temperature. This depends on the stability of the displayed protein. Although it does

not always happen, increasing the temperature often results in increased sensitivity of the fusion protein to *in vivo* proteolysis. The cultivation time is also an important parameter, because it must be long enough for production of a sufficient amount of phages. On the other hand, the level of display, that is, the number of properly folded proteins displayed per phage particle, generally decreases with time of cultivation.

Here are two practical protocols for phage production starting from a stock culture of infected bacteria (stored in 40% glycerol at −80 °C):

- Inoculate 100 µL of infected bacteria into 250 mL of LB medium containing the appropriate antibiotic (tetracycline, if working with fd-DOG1 phages) in a 1-L flask. Incubate with agitation in an orbital shaker at 180 rpm for 20 h at 37 °C or 72 h at 23 °C. This protocol produces large amounts of phages but the level of display is not highly reproducible.
- Inoculate 100 µL of infected bacteria into 25 mL of LB medium containing the appropriate antibiotic and grow overnight at 37 °C. In the morning, centrifuge the culture and resuspend the bacteria in 250 mL of fresh medium. Incubate with agitation in an orbital shaker at 180 rpm for 4 h at 30 °C or 37 °C. This protocol generates phages with more reproducible and generally higher level of display than protocol A, but the amount of phages is lower.

6.3.1.4 Phage-enzyme Purification

Phages are typically purified by successive polyethylene glycol (PEG) precipitations. Nevertheless, the purity is not very high, because some bacterial products coprecipitate with the phages. Moreover, the amount of impurities can vary with the nature of the displayed protein and with time of culture. We suspect that bacterial lysis or periplasmic release can be provoked by phage-enzyme extrusion. Therefore, when high purity is required, a CsCl equilibrium gradient should be performed after PEG precipitations. This also removes the PEG, which probably interferes with the binding of phages to some targets.

6.3.1.4.1 PEG Precipitation

1. Centrifuge a 250-mL bacterial culture at 10 000 rpm for 10 min.
2. Carefully transfer 200 mL of the supernatant to a tube containing 50 mL of a solution containing 20% PEG (w/v)/2.5 M NaCl.
3. Mix thoroughly and incubate for 1 h on ice.
4. Centrifuge at 10 000 rpm for 10 min.
5. Carefully discard the supernatant. Centrifuge again at 10 000 rpm for 1 min and remove the residual liquid.
6. Dissolve the pellet in 20 mL of TBS buffer. Filter on a 1 µm Puradisc 25 GD prefilter unit (Whatman) and then on a 0.45 µm Millex-HV unit (Millipore).
7. Add 5 mL of 20% PEG/2.5 M NaCl, mix well, and incubate 30 min on ice.
8. Repeat steps 4 and 5.
9. Dissolve the pellet in 1 mL TBS. Add 0.02% NaN_3 for long-term storage.

6.3.1.4.2 CsCl Equilibrium Gradient Centrifugation

1. Dissolve 2.5 g of CsCl in 3 mL TE.
2. Add the phage solution (1 mL) and adjust the volume to 5 mL with TE.
3. Centrifuge at 200 000 $\times g$ for 17 h at 15 °C. After centrifugation, the phages appear as a translucent band. PEG appears as a white precipitate below the phage band.
4. Collect the phages by piercing the tube with a needle just below the band and carefully pumping with a syringe. Dialyze twice against TBS. Add 0.02% NaN_3 for long-term storage.

6.3.1.5 Measuring the Phage Titer

1. Because high levels of display can impair phage infection, we recommend treating the phages with 10^{-7} M trypsin for 30 min before measuring the titer. Note that trypsin removes the displayed protein only if a cleavable linker is used or if the protein itself is degraded by trypsin. The stock solution of trypsin (10^{-5} M) should be freshly prepared in 20 mM acetate buffer, pH 3.0.
2. Prepare serial 10× dilutions of the phage solution.
3. Mix 10 µL of these dilutions with 990 µL of a TG1 culture in exponential phase.
4. Incubate at 37 °C without agitation for 30 min and with agitation for another 30 min.
5. Spread 100 µL on petri dishes containing the appropriate antibiotic (tetracycline for fd-DOG1 phages) and incubate overnight at 37 °C.
6. Count the colonies and calculate the phage titer as colony-forming units (cfu).

6.3.1.6 Measuring Phage Concentration

The phage concentration is simply obtained by measuring the absorbance at 265 nm and using the appropriate extinction coefficient. Because the phage size varies with the length of the inserted gene, the coefficient is proportional to the genome size. For a 10-kb phage, the extinction coefficient is $8.4 \times 10^7\ M^{-1}cm^{-1}$. Note that the phage concentration is generally 20–50 times higher than the phage titer.

6.3.1.7 Measuring the Activity of a Phage-enzyme

A solution of phage-enzyme can be used like an enzyme solution for measuring kinetic parameters such as k_{cat} and K_M. We usually observed that phage-displayed enzymes behave essentially like free enzymes in solution, although interference is always possible, especially with multiple display. Note that the k_{cat} is the turnover rate of the phage and not of the enzyme, because the level of display is generally just an evaluation (see below).

6.3.1.8 Evaluating the Level of Display

The level of display is the average number of enzymes displayed per phage particle. For a wild-type enzyme whose activity is not affected by the phage environment, the level of display is evaluated by dividing the k_{cat} of the phage by the k_{cat} of the free enzyme. Otherwise, it can be evaluated by Western blot or, when possible, by active-site labeling.

6.3.1.8.1 Western Blot

The protocol involves a classical SDS-PAGE (10% polyacrylamide) run, followed by transfer onto a Western blot membrane and immunodetection with an anti-pIII antibody. Nevertheless, special care must be taken during sample preparation, because phages are very stable and difficult to denature. The protocol is similar to typical SDS–PAGE sample preparation, except that β-mercaptoethanol should be replaced by fresh dithiothreitol (DTT, 5 mM final concentration), and the samples should be boiled in a water bath for at least 15 min. Moreover, because the pIII-fusion protein is a minor component of the virion, a large amount of phages should be loaded onto the gel, typically around 10^{12} phages per lane.

The level of enzyme display is evaluated by comparing the relative intensities of the bands corresponding to the enzyme-pIII fusion protein and the pIII protein alone. For example, if equivalent band intensities are observed, it means that approximately 50% of the pIII are expressed as fusion proteins. Hence, considering 3–5 copies of pIII per phage particle, the average level of display is evaluated as 1.5–2.5 enzymes per phage. A rough quantitative value is obtained by using a fluorescent detection protocol and a fluorescence imaging equipment (e. g., FluorImager 595 and ECL detection kit, Amersham Biosciences, Cardiff, UK) for band quantitation.

6.3.1.8.2 Active-site Labeling

Whenever active site labeling of the displayed enzyme is possible with fluorescent or radioactive compounds, the level of display should also be evaluated by this method. Although the detailed protocol depends on the displayed enzyme, it must involve the following steps: (1) labeling the phages, (2) removing the excess label by two PEG precipitations or by dialysis, and (3) comparing the fluorescence or the radioactivity of the phages with standards. The phage concentration should not be higher than 10 nM, to avoid solubility problems.

6.3.2 Library Construction

6.3.2.1 Type of Library

Although the choice is somewhat arbitrary, the nature of the library must be chosen by evaluating the chances of finding interesting mutants in it. The four major methods for creating phage-display libraries are error-prone PCR (see Chapter 2), DNA

shuffling (see Chapter 4), PCR with degenerate primers, and cloning of degenerate oligonucleotides. All these methods are also largely described in the literature (for another laboratory manual, see [21]). Hence, this section does not focus on detailed protocols but rather on practical tricks.

6.3.2.1.1 Error-prone PCR

Error-prone PCR consists in performing PCR amplification under conditions in which polymerase fidelity is decreased [22]. Unbalanced deoxynucleotide triphosphate concentrations are used, and manganese ions are added. It is important to note that not all possible mutations (on the amino acid level) are represented in such a library, because only single-nucleotide substitutions are obtained. Moreover, some mutations are favored over others. The average number of mutations per gene depends on the number of PCR cycles. Although this number should be predictable, some parameters such as the amount of starting phage RF DNA, which is always contaminated with single-stranded DNA, and the rate of exponential amplification are difficult to control. Therefore, before creating the library, it is recommended to sequence a few clones to verify that the average number of mutations is close to the expected number. It is also recommended to use primers relatively far from the chosen restriction sites, to facilitate the subsequent restrictions and easily monitor these restrictions by agarose gel electrophoresis.

6.3.2.1.2 DNA Shuffling

W.P.C. Stemmer introduced the DNA shuffling technology in 1995 [23] and, since then, it has become a very popular method for creating libraries. It requires starting with a family of highly homologous genes, which is not always available. The method can also be very powerful for creating a second-generation library from an ensemble of clones that have been selected from an initial library. The major advantage of the technique is that shuffling should allow the combination of favorable mutations and the removal of deleterious ones. Nevertheless, the method results also in the introduction of point mutations, and successive rounds of selection and shuffling should therefore be avoided.

6.3.2.1.3 PCR with Degenerate Primers

Degenerate oligonucleotides can be introduced into a gene by using them as PCR primers. These primers generally comprise a complementary sequence of about 15 nucleotides for hybridization that is followed by the degenerate sequence and a sequence containing either a restriction site for direct cloning or one that is designed for gene assembly by an overlap extension PCR. Obviously, PCR with a degenerate primer produces a fraction of DNA fragments with incomplete hybridization that cannot be restricted and cloned. This problem can be partly overcome by stopping the PCR during the exponential amplification phase, to avoid successive denatura-

tion and hybridization steps without replication. Therefore, the optimal number of PCR cycles should be determined prior to the library synthesis by PCR.

6.3.2.1.4 Cloning Degenerate Oligonucleotides

In this method, a degenerate oligonucleotide is synthesized with two flanking sequences containing specific restriction sites. A small oligonucleotide, complementary to the 3′ region, is then annealed and a double-stranded cassette is generated by polymerization with T4 DNA polymerase or Klenow polymerase. The degenerate cassette is then restricted and cloned into the phage-display vector. Hence, specific restriction sites must be present in the vector. While designing the oligonucleotide, you should keep in mind that the flanking sequences next to the restriction sites must be of sufficient length for efficient restriction. Although the number of nucleotides in the 'overhang' varies with the restriction enzyme, we recommend at least 10 nucleotides. It might also be interesting to take advantage of enzymes that cut outside their recognition sequence, such as *Bbs*I. With these restriction sites, it is possible to eliminate the recognition sequences during cloning both in the vector and in the cassette. Moreover, restriction with these enzymes often generates nonpalindromic cohesive ends, which should significantly increase the ligation efficiency.

6.3.2.2 Library Diversity

The diversity is the number of individual clones that are present in the library. The chances of finding interesting clones in a library increases with increasing diversity. Hence, when constructing a phage library, achieving the highest diversity is a major concern. Several methods are available for constructing libraries, but the last step always involves ligating a collection of DNA fragments into a phage or a phagemid vector, followed by transformation. The transformation efficiency generally determines the diversity and therefore, must be optimized. Here are some practical tricks concerning the final ligation and transformation steps:

- Use large amounts of well purified, restricted vector (50 µg in 300 µL of a ligation reaction). We recommend preparing the vector with a Qiagen maxipreparation kit and purifying it further by CsCl gradient centrifugation.
- Purify the collection of insert DNA fragments by polyacrylamide gel electrophoresis.
- Use a ratio of vector to fragment of 1/3 for the ligation.
- Purify the ligation mix as thoroughly as possible. For efficient transformation, it is important to remove all the ions by passing through a desalting column or by dialysis. Concentrate the DNA to approximately 1 µg µL^{-1}.
- Make a transformation test before creating the library, by electroporating 100 µL of competent cells with 1 µL of the ligation mix, and compare the transformation efficiency with that of a standard plasmid like pUC18. For reaching high library sizes, each transformation should yield between 10^6 and 10^7 transformants. Repeat electroporations to reach the desired diversity.

6.3.2.3 Library Production

Transformed bacteria must be grown to produce the phage library. At this stage, *in vivo* selection processes may favor some clones over others. To reduce these biases, it is recommended to avoid liquid culture for the first library production. Hence, the library should be grown on large petri dishes ($30 \times 30\,cm^2$) containing agar medium. No more than 10^8 individual transformants are grown per petri dish. A typical protocol for preparing the first generation library is as follows:

1. After electroporation, add 900 µL of LB to 100 µL cells. Incubate 1 h at 37 °C.
2. Take an aliquot of 10 µL and add 990 µL of LB medium. On small petri dishes, plate 100 µL of serial 10× dilutions of these cells for measuring the library diversity.
3. Spread the electroporation mix on a large petri dish containing LB agar and the appropriate antibiotic (tetracycline for fd-DOG1 phages). Incubate overnight at 37 °C or 72 h at 23 °C.
4. Recover the bacteria and the phages by pouring 30 mL TBS onto the agar and resuspending the bacteria. Repeat this step 3 times per petri dish.
5. Spin down the bacteria at 10 000 rpm for 10 min and recover the phages in the supernatant.
6. Purify the phages by PEG precipitations (see Section 6.3.1.4.1) and store the library at 4 °C.
7. The bacteria should also be resuspended in LB containing 40% glycerol and stored at −80 °C.

Every time the library must be produced for selection, it should be prepared from the first-generation library by infection.

1. Take an aliquot of phages containing at least 100 times more phages than the number of variants expected to be present in the library (= library diversity).
2. Add 10^{-7} M trypsin for removing the displayed enzymes (Section 6.3.1.5) and for preventing infection bias. Incubate 30 min at room temperature. Remove the trypsin by precipitating the phages with PEG.
3. Dissolve the phages in 1 mL TBS and infect a TG1 culture in exponential phase. The culture should contain at least 10 times more bacteria than the library diversity; 1 OD_{600} corresponds to approximately 10^8 bacteria mL^{-1}.
4. Incubate 1 h at 37 °C without agitation.
5. Mix the cells and take an aliquot for phage titering. The titer should be at least 10 times higher than the library diversity.
6. Grow the bacteria at 37 °C for 4 h or overnight at 23 °C with agitation. Growing at a lower temperature improves the display of unstable enzymes.
7. Purify the phages as described (Section 6.3.1.4).

6.3.3 Selection

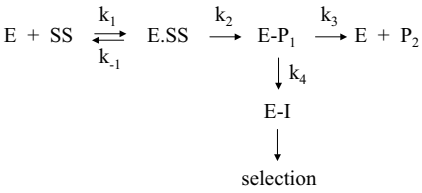
Whenever possible, a model selection should be optimized before starting selections with a library. In this experiment, a mixture of active and inactive phage-enzymes is used as a model library for one round of selection. The phage mixture is analyzed before and after selection, yielding numbers that serve for the calculation of an enrichment factor (EF):

$$\mathrm{EF} = \frac{(A_{\mathrm{out}}/I_{\mathrm{out}})}{(A_{\mathrm{in}}/I_{\mathrm{in}})}$$

Here, A and I represent the fractions of active and inactive phages, respectively, after (out) or before (in) the selection. A selection strategy is considered efficient if the enrichment factor is higher than 50.

6.3.3.1 Selection with Suicide Substrate

This strategy is based on selecting the phages displaying active enzymes by labeling them with a biotinylated suicide substrate and capturing the labeled phages with immobilized streptavidin. The strategy was initially developed for selecting enzymes that feature a covalent intermediate in their mechanism of action [5]. For these enzymes, a general suicide mechanism can be schematized as follows:



Scheme 6.1

6.3.3.1.1 Labeling

As illustrated in Scheme 6.1, once the covalent intermediate is formed, the complex can either follow a normal catalytic cycle or go through a suicide event leading to the irreversible labeling that is necessary for selection. The suicide inhibition efficiency depends on the ratio k_4/k_3. This ratio depends on the nature of the suicide substrate and of the enzyme. Therefore, a large excess of suicide substrate as compared to the displayed enzyme is recommended for selection experiments.

Because the labeling event occurs before the last step of the turnover, and because it even competes with this last step, there is selective pressure for enzymes that feature very low k_3 rates, that is, enzymes with very low turnover. To overcome this problem, active sites that do not turn over rapidly can be blocked by an initial incubation with a normal substrate prior to labeling with the biotinylated suicide substrate. Hence, special care must be given to the kinetic control of the labeling step.

The following protocol is described for the selection of phage-displayed serine β-lactamase with a biotinylated penam-sulfone [5] suicide substrate. For other activities, the concentrations of substrate and suicide substrate and the times of reaction should probably be adjusted.

1. Just prior to use, prepare 1 mM stock solutions of substrate and biotinylated suicide substrate in TBS buffer. If the solubility is too low, dissolve in pure DMSO. Note that a small amount of water in the DMSO could degrade the substrates rapidly, because the activity of water is high in DMSO. Therefore, using freshly dried DMSO is recommended.
2. In a final volume of 1 mL phosphate buffer (50 mM, pH 7.0), mix 10^{12} phages with 10^{-5} M of substrate and incubate 10 min at room temperature to block all the active sites that are inactive.
3. Add 400 µL of PEG (20%)/NaCl (2.5 M), vortex a few seconds, and centrifuge 5 min at 14 000 rpm. Discard the supernatant containing the excess substrate and dissolve the phage pellet in 1 mL phosphate buffer.
4. Add the biotinylated suicide substrate to a final concentration of 10^{-5} M and incubate 20 min at room temperature.
5. Eliminate the excess suicide substrate by two PEG precipitations as described in step 3.
6. Take an aliquot of 10 µL for measuring the 'input' phage titer and proceed immediately to the capture with the remaining 990 µL.

6.3.3.1.2 Capture and Elution

1. In a microtube, add 1 mg of M-270 streptavidin-coated Dynabeads to 1 mL TBS and place the tube on a magnet to discard the supernatant. Off the magnet, resuspend the beads in 1 mL of MTBS for blocking nonspecific sites. Place the microtube on a rotating wheel for 1 h at room temperature. The rotation is necessary for keeping the beads in suspension and should be slow.
2. Remove the MTBS from the magnet. Wash the beads with 1 mL TBS and resuspend the beads in the phage solution supplemented with 1% bovine serum albumin (BSA).
3. Place the microtube on a rotating wheel for 4 h at room temperature or overnight at 4 °C.
4. On the magnet, discard the supernatant containing unbound phages. Eventually take an aliquot for measuring the phage titer.
5. Wash the beads 5 times with 1 mL TTBS and 1 time with TBS. A single washing consists in resuspending the beads off the magnet and discarding the solution on the magnet.
6. If the connecting linker or the displayed enzyme is susceptible to proteolytic cleavage, resuspend the beads in 1 mL TBS containing 10^{-7} M trypsin or 5 units of factor Xa. If the suicide substrate contains a disulfide bridge, elution can also be performed by resuspending the beads in 1 mL TBS containing 10 mM DTT. In both cases, incubate on the rotating wheel for 1 h.

7. Recover the phages in the supernatant and add them to 50 mL of an exponential-phase culture of TG1 in LB medium. Mix well and incubate at 37 °C without agitation for 30 min and with agitation for 30 min.
8. Take 300 µL for measuring the 'output' phage titer and transfer the rest of the culture to a 1-L flask containing 200 mL of LB medium with the appropriate antibiotic. Proceed as described in Section 6.3.1.3.

6.3.3.1.3 Successive Selection Rounds

The selection efficiency is evaluated by the ratio of the 'output' over 'input' titers. After each selection, the phages are amplified and can be subjected to a new selection round. If some clones of the library are effectively selected, the ratio should increase from around 10^{-5} (background level) to around 10^{-2}. Typically, the ratio reaches a plateau after about 4 to 8 selection rounds – depending on the starting diversity and on the power of the selection itself.

It can be interesting to increase the selection pressure from one selection round to the next by doubling the number of washes, by decreasing the suicide substrate concentration, or the time of incubation by a factor of 10. This could lead to the selection of the most efficient catalysts.

At least 20 clones resulting from the last round before the plateau is reached should be sequenced for evaluating the diversity after selection. Depending on the complexity of the activity assay, as many clones as possible should also be screened for activity. Monoclonal preparations of phage-enzymes should be assayed first and, if the activity is too low, soluble overexpressed enzymes should be produced for reaching higher concentrations.

6.3.3.2 Other Selection Strategies

As mentioned in the introduction, several other strategies have been developed for selecting enzymes on phages. Because most of the protocols are similar, this section focuses only on the major technical and practical differences. The selection with transition-state analogs, which is a simple affinity selection, is not described here.

6.3.3.2.1 Selection by Using a Suicide Leaving Group

This strategy is a slight variant of the one previously described. The nature of the selecting substrate is different inasmuch as it contains a suicide leaving group [7,8]. Here, enzyme turnover releases a reactive species that ultimately reacts with a proximal residue and labels the phage. The advantage is that selection does not require a stable intermediate. Nevertheless, these substrates are very poor suicide substrates because of the rapid diffusion of the leaving group. To overcome this problem, it is necessary to perform the reaction with a substrate that has been immobilized via the leaving group side. This solves the problem of rapid diffusion and, because the phage-enzyme is not diffusing rapidly in solution either, the labeling is efficient enough for selection.

Working with an immobilized substrate, however, can be a source of problems that are difficult to identify. Because of the slow diffusion of the phages, the time of reaction should be greatly increased. Control of some parameters, such as interference by the support or the density of immobilization, is difficult. Moreover, substrate recognition by the enzyme can be significantly impaired.

6.3.3.2.2 Selection by Catalytic Elution

In this strategy, phages are affinity-captured on immobilized substrate under conditions in which the enzyme is inactive, for example, in the absence of an essential metal ion or a cofactor. Active phage-enzymes are then eluted by triggering catalysis by addition of the metal ion or cofactor, taking advantage of the lower affinity for the product than for the substrate.

Meanwhile, because enzymes do not generally have high affinity for their substrates, the initial affinity capture may be problematic. It is therefore advised to generate phages with a high level of display, that is, more than one enzyme per phage, to take advantage of possible avidity effects. It is also important to remove the displayed enzymes by proteolysis after the elution, because multivalent display can impair phage infection. Besides these aspects, the selection is essentially like a classical affinity selection.

6.3.3.2.3 Selection by Product Labeling

As shown in Figure 6.1, the protocol starts with labeling the phages with the substrate. Several approaches have been used for this labeling step [11, 13, 14]. The active enzymes are then turning over the substrate into product by intraphage catalysis. Finally, the product-labeled phages are selected by classical affinity capture.

In this strategy, it is very important to avoid interphage catalysis. Therefore, the phage concentration should be kept low (lower than 10^{-9} M) and phages should not be precipitated with PEG. Because removal of excess label is generally required before affinity capture, it should be done by phage dialysis or size-exclusion chromatography.

6.4 Troubleshooting

6.4.1 Phage Titers Are Not Reproducible

Phages are sticky to themselves and to solid supports. When they are concentrated, they can form soluble aggregates that dissociate relatively slowly. Therefore, it is recommended to thoroughly vortex phage solutions before infection. Because they also stick to micropipet tips, it is recommended to change tips when performing serial dilutions. Finally, when phages are highly diluted, a time-dependent loss of

infection can result from their binding to the vessel walls. It is therefore recommended to use silanized microtubes or to add 1% BSA to the solution or to avoid keeping highly diluted solutions for long periods of time.

6.4.2 Phage-enzymes Degrade with Time

Degradation could be due to the presence of proteases or to inherent low enzyme stability. For prevention, a cocktail of protease inhibitors can be added (e. g., Complete tabs, Roche, Penzberg, Germany). Always use freshly prepared phage solutions when performing selection from libraries.

6.4.3 Phages Are Not Genetically Stable

Genetic instability could be due to low toxicity of the fusion protein or to recombination with homologous *E. coli* genes. Use a *recA* strain such as JM109 to reduce recombination. For the toxicity problem, use a phagemid vector such as pHDi.Ex [17], which allows tight control of the fusion protein expression. Hence, the expression can be completely repressed during a culture in which a large bacterial mass is produced. Then induction is triggered simultaneously with phage helper infection, and phages are produced over a short period of time.

6.4.4 The 'out/in' Ratio Does Not Increase with Selection Rounds

A constant 'out/in' ratio could mean that no clones are being selected. For some strategies that require low-affinity capture for selection, the level of specifically captured phages might always be below the background level. It is therefore worth analyzing the selected phages, because effective enrichment may have occurred.

6.5 Major Applications

Applications of the selection of phage-displayed enzymes are obvious in the field of enzyme engineering, ranging from the improvement of existing enzymes to the creation of new catalytic activities. In the past, the selection of phage-displayed enzymes was essentially performed with model mixtures to demonstrate the feasibility and the potentials of the technology. The maturity of the technological developments is now sufficient for going into selections from libraries.

So far, only a few groups have reported the application of this technology to libraries. As yet, suicide substrates have been applied for the selection of β-lactamase activity within libraries of mutants containing penicillin-binding proteins [6] and to

the selection of subtilisin variants with altered substrate specificity [9]. The suicide leaving group approach has been applied for the selection of catalytic antibodies endowed with glycosidase and phosphatase activity [7, 8].

Directed enzyme evolution using the selection of phage-displayed enzymes is a powerful tool that represents an attractive alternative to the popular high-throughput screening technologies.

References

1. G.P. Smith, *Science* **1985**, *228*, 1315–1317.
2. G.P. Smith, V.A. Petrenko, *Chem. Rev.* **1997**, *97*, 391–410.
3. A. Fernandez-Gacio, M. Uguen, J. Fastrez, *Trends Biotechnol.* **2003**, *21*, 408–414.
4. M. Widersten, L.O. Hansson, L. Tronstad, B. Mannervik, *Methods Enzymol.* **2000**, *328*, 389–404.
5. P. Soumillion, L. Jespers, M. Bouchet, J. Marchand-Brynaert, G. Winter, J. Fastrez, *J. Mol. Biol.* **1994**, *237*, 415–422.
6. S. Vanwetswinkel, B. Avalle, J. Fastrez, *J. Mol. Biol.* **2000**, *295*, 527–540.
7. K.D. Janda, L.C. Lo, C.H. Lo, M.M. Sim, R. Wang, C.H. Wong, R.A. Lerner, *Science* **1997**, *275*, 945–948.
8. S. Cesaro-Tadic, D. Lagos, A. Honegger, J.H. Rickard, L.J. Partridge, G.M. Blackburn, A. Pluckthun, *Nat. Biotechnol.* **2003**, *21*, 679–685.
9. D. Legendre, N. Laraki, T. Graslund, M.E. Bjornvad, M. Bouchet, P.A. Nygren, T.V. Borchert, J. Fastrez, *J. Mol. Biol.* **2000**, *296*, 87–102.
10. S. Danielsen, M. Eklund, H.J. Deussen, T. Graslund, P.A. Nygren, T.V. Borchert, *Gene* **2001,** *272*, 267–274.
11. H. Pedersen, S. Holder, D.P. Sutherlin, U. Schwitter, D.S. King, P.G. Schultz, *Proc. Natl Acad. Sci. USA.* **1998**, *95*, 10523–10528.
12. I. Ponsard, M. Galleni, P. Soumillion, J. Fastrez, *Chembiochem* **2001**, *2*, 253–259.
13. S. Demartis, A. Huber, F. Viti, L. Lozzi, L. Giovannoni, P. Neri, G. Winter, D. Neri, *J. Mol. Biol.* **1999**, *286*, 617–633.
14. J.L. Jestin, P. Kristensen, G. Winter, *Angew. Chem. Int. Ed.* **1999**, *38*, 1124–1127.
15. C. Heinis, A. Huber, S. Demartis, J. Bertschinger, S. Melkko, L. Lozzi, P. Neri, D. Neri, *Protein Eng*. **2001**, *14*, 1043–1052.
16. T. Clackson, H.R. Hoogenboom, A.D. Griffiths, G. Winter, *Nature* **1991**, *352*, 624–628.
17. B. Heyd, F. Pecorari, B. Collinet, E. Adjadj, M. Desmadril, P. Minard, *Biochemistry*. **2003**, *42*, 5674–5683.
18. B.K. Kay, J. Winter, J. McCafferty, *Phage Display of Peptides and Proteins: A Laboratory Manual*, Academic Press, San Diego, CA **1996**.
19. P.M. O'Brien, R. Aitken, *Antibody Phage Display: Methods and Protocols*, Humana Press, Totowa, NJ **2001**.
20. C.F. III. Barbas, J.K. Scott, G. Silverman, D.R. Burton, *Phage Display: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Woodbury, NY **2001**.
21. J. Braman, *In Vitro Mutagenesis Protocols, second edition*, Humana Press, Totowa, NJ **2001**.
22. R.C. Cadwell, G.F. Joyce, *PCR Methods Appl.* **1994**, *3*, S136–S140.
23. W.P.C. Stemmer, *Nature* **1994**, 370, 389–391.

7 Selection of Aptamers

Heiko Fickert, Heike Betat, and Ulrich Hahn

7.1 Introduction

Aptamers are nucleic acids which exhibit a defined structure due to their nucleotide sequence and therefore, are able to specifically bind selected targets [1] (*aptus* [lat.] = fitting, sticking to). Aptamers and likewise, ribozymes [2] and deoxyribozymes [3] are selected *in vitro* by screening nucleic acid libraries. Here we describe in detail the selection of aptamers by a process called SELEX (Systematic Evolution of Ligands by EXponential enrichment) [4].

Usually, one starts with a nucleic acid library comprising 10^{14} to 10^{16} individual molecules [5]. This library size is assumed to be sufficient to contain sequences with the desired property [6]. Additional mutations may be introduced into selected nucleic acid variants by repeating the SELEX cycle, thereby increasing the number of screened nucleic acids with different sequences.

The general course of a SELEX experiment is illustrated in Figure 7.1 using the isolation of a target-binding RNA, that is, an RNA-aptamer, via affinity chromatography as an example. Starting with a pool of chemically synthesized single-stranded (ss) DNA oligonucleotides, double-stranded (ds) DNA variants are generated by polymerase chain reaction (PCR) (step 1). The DNA templates contain a region of 20–60 randomized bases that are flanked by constant regions (primer binding sites) necessary for amplification of the selected molecules by PCR and reverse transcription (RT). One of the PCR primers must contain the T7 promoter consensus sequence that is required for the synthesis of the RNA pool by transcription using T7 RNA polymerase (2). For selection by affinity chromatography, the target molecule must be immobilized on a suitable matrix. To prevent enrichment of RNA molecules that bind to the matrix devoid of a binding target, a preselection step with underivatized matrix is carried out. RNA that is either not bound or just weakly bound is eluted from the preselection column (3) and is subsequently used in the main selection step in which the pre-selected RNA pool is incubated with the target-modified matrix (4). RNA that is either not bound or just weakly bound during the selection step is washed away (5), and the bound RNA is eluted (6). Afterwards, the selected RNA molecules are reverse-transcribed to yield cDNA (7), which is then amplified by PCR to create the DNA pool for another round of the SELEX procedure (8).

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst
Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-30799-0

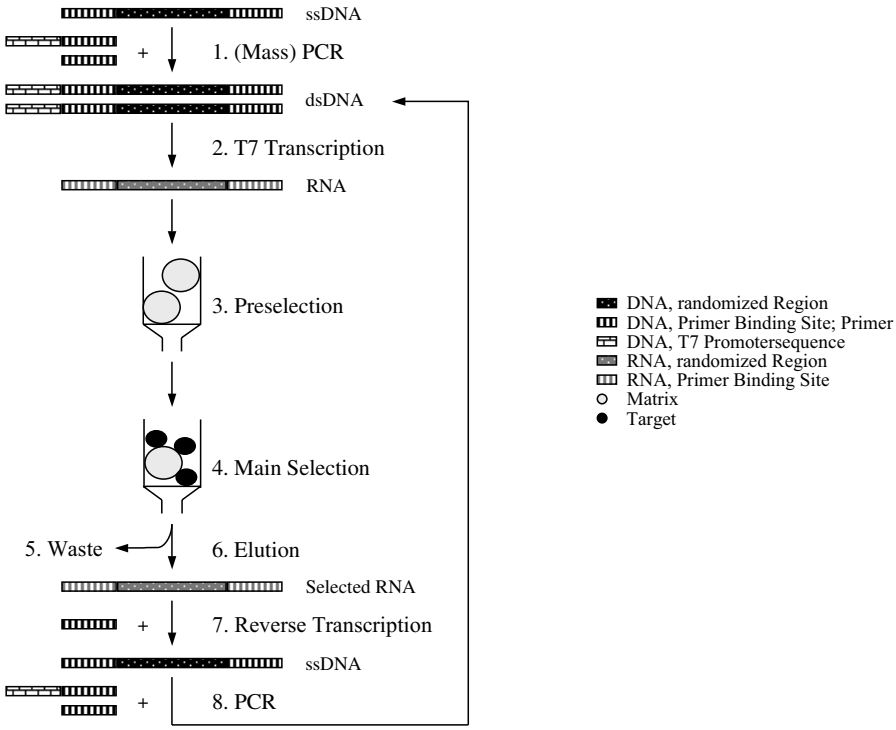


Fig. 7.1. Selection of RNA aptamers. For detailed explanation, see Section 7.1.

Usually, a minimum of five repetitions of steps 2–8 is required to yield an RNA pool that predominantly consists of the best target-binding RNA molecules. Individual aptamers are then isolated from this pool by cloning, and identified by sequencing.

7.2 Materials

Materials differ significantly with the kind of target and selection procedure. Here we list materials used for the example selection of moenomycin A-specific 2′-modified RNA aptamers (Section 7.3.6).

7.2.1 Immobilization of Target Molecules

- Activated thiol Sepharose 4B (Pharmacia, Freiburg, Germany)
- Moenomycin-S-S-pyridone (gift from Peter Welzel, Leipzig)

- Immobilization buffer: 500 mM NaCl, 1 mM EDTA, 10 mM sodium acetate, pH 5.0
- Selection buffer: 20 mM HEPES, pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$

7.2.2 PCR

- DNA pool (synthesized using a Millipore Expedite DNA synthesizer): 5′-C TAT AGG GAG AGA CAA GCT TGG GTC - $\mathbf{N_{40}}$ - AGA AGA GAA AGA GAA GTT AAT TAA GGA TCC TCA C-3′
- Primer A (purchased from Metabion, Munich, Germany): 5′-TCT AAT ACG ACT CAC TAT AGG GAG AGA CAA GCT TGG GTC-3′
- Primer B (purchased from Metabion, Munich, Germany): 5′-CTG AGG ATC CTT AAT TAA CTT CTC TTT CTC-3′
- dNTPs (MBI Fermentas, Vilnius, Lithuania)
- 10× PCR buffer: 25 mM TAPS, pH 8.3, 50 mM KCl, 17 mM $MgCl_2$, 1 mM β-mercaptoethanol
- DapGoldStar DNA polymerase (Eurogentec, Liege, Belgium)

7.2.3 *In vitro* Transcription

- ATP, GTP (Boehringer Mannheim, Mannheim, Germany)
- 2′-NH_2-CTP, 2′-NH_2-UTP (Amersham Biosciences, Freiburg, Germany)
- [α-^{32}P]-GTP (800 Ci mmol^{-1}) (PerkinElmer Life Sciences, Rodgau-Jügesheim, Germany)
- 5× transcription buffer: 400 mM HEPES-KOH, pH 8.0, 60 mM $MgCl_2$, 10 mM spermidine, 25 mM dithiothreitol
- T7 RNA polymerase (Stratagene, Heidelberg, Germany)

7.2.4 RNA Purification

- DNase I (Boehringer Mannheim, Mannheim, Germany)
- Polyacrylamide gel electrophoresis equipment

7.2.5 Selection of Aptamers

- Selection buffer: 20 mM HEPES, pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$
- RNA
- Thiol sepharose
- Moenomycin sepharose

7.2.6 Reverse Transcription

- Super Script™ II reverse transcriptase (Gibco BRL, Eggenstein, Germany)
- 5× reverse transcription buffer (Gibco BRL, Eggenstein, Germany)

7.3 Protocols

In this chapter, we introduce the essential steps that must be employed for the selection of (i) RNA aptamers, (ii) RNA aptamers modified at the 2′-position, and (iii) DNA aptamers, and we present experimental approaches to their realization. To cope with the variety of different selection problems, which depend strongly on the uniqueness of the chosen target and the intended application of the aptamer, we describe alternative procedures in some instances.

For example, aptamers specific for small molecules like amino acids [7] usually bind their target molecules with lower affinities than aptamers specific for larger molecules like proteins [8]. In consequence, different targets must be applied at different concentrations – depending on the K_D value of the best-binding aptamer. For example, if the concentration of a certain target is lower than the K_D of the best-binding aptamers in the pool, just a small portion of these aptamers will bind. Thus, it may be impossible to enrich aptamers if the number of nucleic acid molecules is too small and/or the washing procedure is too stringent. If, on the other hand, the target concentration is too high, weakly binding aptamers will be selected because discrimination between weakly and tightly binding RNA molecules will not occur.

Also, the selection conditions may vary substantially depending on target properties and the intended applications of the aptamers. For example, aptamers that are intended for intracellular applications (intramers) [9] are usually selected under physiological conditions. For the selection of highly specific aptamers, a preselection step should be included that employs a 'target-related' molecule that should not be recognized by the aptamers [10]. Aptamers with low k_{off} rates can be isolated if the affinity column with the bound aptamers is washed with selection buffer containing free target molecules prior to the final elution. Details of these different approaches are described below.

7.3.1 Selection of RNA Aptamers

Figure 7.2 shows the individual steps of the isolation of RNA aptamers. The steps of one selection cycle are highlighted in gray. Usually, it takes 2 or 3 days to complete one cycle. Robots, however, can perform such an experiment in just a few hours [11].

After 5 or, up to > 15 rounds of selection, single aptamers can be isolated and identified by cloning and sequencing. The characterization of aptamers can be

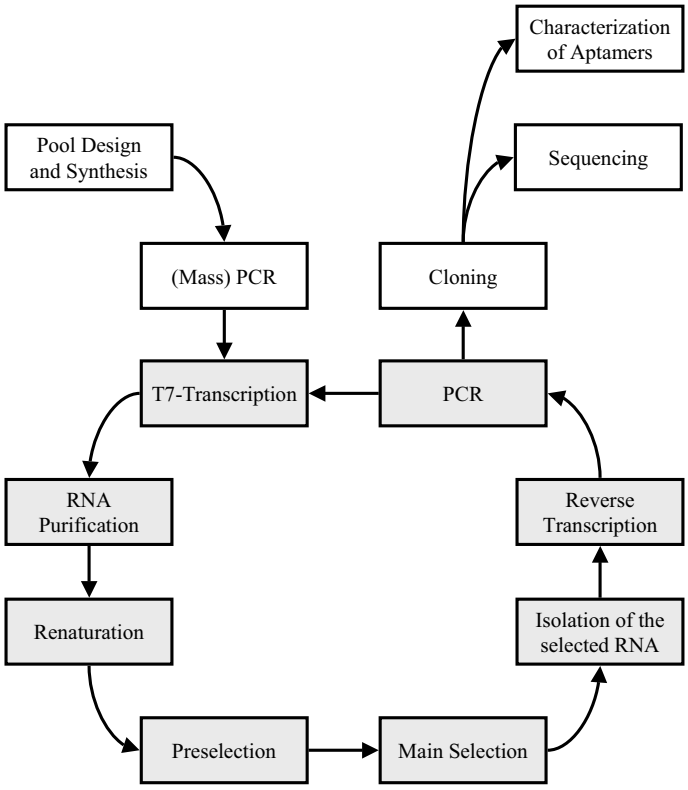


Fig. 7.2. *In vitro* selection of an RNA aptamer. Steps of one cycle are given in gray.

achieved with different methods. The required time depends on the methods used and the number of RNAs to be characterized.

7.3.1.1 Starting Pool Design

In almost every published SELEX protocol, the experiment starts with a pool of chemically synthesized ssDNA. The DNA usually consists of a central randomized region of 20–60 nucleotides that is flanked by two constant regions that are necessary for primer binding.

The randomized region is obtained by using a mixture of all four bases in each synthesis step. In most cases, the four building blocks are mixed with balanced stoichiometry (IUB mix code: 'N'). However, pools may also be synthesized for the selection of aptamers that contain only three different bases in the randomized region [12] or unequal frequencies of the four bases [13].

Such ssDNA pools are available from various commercial suppliers of oligonucleotides. The main tasks in designing starting pools concern the definition of primer

binding sites, which may include recognition sites of restriction endonucleases for cloning [14], and the definition of the randomized region with respect to length, frequency of bases, and possible constant positions.

In some cases, pools are made of genomic DNA [15] or by combining several synthetic pools to form longer randomized regions [16]. Concerning the length of the randomized region, there is no defined optimum; however, most experiments that yielded good aptamers were started with pools of 40–60 randomized nucleotide positions. Thus, these values can be regarded as sufficient [17].

7.3.1.2 DNA Template Amplification

Amplification of the DNA template is achieved by PCR (for details, see [14]). The size and amount of the PCR product should be routinely checked by agarose gel electrophoresis. An insufficient amount of product may be improved by increasing the number of PCR cycles. However, care should be taken not to overdo the cycling, because this leads to a loss of dsDNA: after several PCR cycles the amount of free primers decreases dramatically so that only a fraction of the DNA template strands anneal with a primer and yield new dsDNA. The high complexity of the pool causes that ssDNA is not able to rehybridize with the complementary strand.

Furthermore, the ssDNA cannot be transcribed by T7 RNA polymerase, which recognizes a double-stranded promoter region of at least 17 base pairs. However, if agarose gel electrophoresis reveals the presence of ssDNA (which runs more slowly than the corresponding dsDNA), these molecules can be converted into dsDNA by adding further primers to the PCR mixture and using one extra PCR cycle.

These recommendations also hold for *DNA-Template Amplification after Reverse Transcription* which is described in Section 7.3.1.6.

DNA-Template Amplification of the Starting Pool

The ssDNA pool first has to be amplified by PCR in the presence of the two primers to obtain dsDNA. By this step, multiple pool copies can be generated which can be used in several SELEX experiments.

Note that the different variants found in the starting pool are amplified with different efficiencies during PCR. In most cases, only a fraction (about 30% of the chemically synthesized ssDNA) is amplifiable by PCR [18]. As a consequence, the distribution of variants differs substantially from that in the starting pool if the number of cycles is too high. Therefore, we recommend the application of only 4–6 PCR cycles. After PCR, the dsDNA product can be purified by ethanol precipitation. Routinely, the library should be dissolved and stored in ddH_2O.

7.3.1.3 *In vitro* Transcription

The dsDNA pool is transcribed into the corresponding RNA pool with T7 RNA polymerase [19] during 2–4 hours of incubation (samples may also be left overnight).

Other RNA polymerases, like SP6 or T3, can also be used, provided that the corresponding promoter has been introduced. By using [α-^{32}P]-nucleotides the RNA can be labeled radioactively for easy determination of the amount of selected RNA. Thus, the progress of the SELEX process can be monitored.

7.3.1.4 RNA Purification

The RNA purification step is necessary for removal of DNA, unincorporated nucleotides, and transcription byproducts. DNA must be removed because it can obstruct subsequent reverse transcription and PCR (Section 7.3.1.6). This can be achieved by DNase I digestion and/or by denaturing polyacrylamide gel electrophoresis (PAGE, 6%–10 % polyacrylamide, 7 M urea) [14]. Unincorporated nucleotides and transcription byproducts are best removed by denaturing PAGE.

DNase I Digestion

DNA is usually degraded with DNase I, which is simply added to the *in vitro* transcription mixture. We highly recommend checking this step, because we have observed that it can be necessary to optimize DNase I digestion to remove even the smallest trace of amplifiable DNA. Therefore, the overall nucleic acid concentration in the digestion mixture should not exceed 100 µg mL^{-1}, and the DNA concentration should be less than 10 µg mL^{-1}. NEBuffer 1 (New England BioLabs, Frankfurt am Main, Germany) turned out to be optimal for us. Of course, a buffer change requires an ethanol precipitation in between.

Denaturing Polyacrylamide Gel Electrophoresis (Denaturing PAGE)

First, the RNA must be collected from the transcription mixture by precipitation. We use ammonium acetate in the ethanol precipitation step because this is the quickest method. Coprecipitated NH_4^+ ions, which interfere with further enzymatic reactions, are removed in the next purification step. The RNA pellet is dissolved in an appropriate loading buffer for denaturing PAGE (7 M urea, 50 mM EDTA), incubated for 5 min at 65 °C, and then immediately loaded on the gel. The gel should be preequilibrated and preheated by preelectrophoresis for 15 min.

After electrophoresis (1 W cm^{-1} gel width, for about 1.5 hours; alternatively a 20 × 20 cm gel should be run at 500 V), the gel piece containing the RNA of correct size is cut out and the RNA is eluted (see below). If the amount of RNA is large enough, the sample can be detected by UV shadowing [20]; if not, the gel must be stained with toluidine blue O, Sybr Green II, or ethidium bromide. In any case, care should be taken to completely remove the dye from the RNA after staining (for example, by ion-exchange chromatography or extraction with organic solvents [14]).

Elution of the RNA from the gel pieces can be accomplished by diffusion ('crush-and-soak method') or by electroelution (for example, electroelution into dialysis bags) [14]. Afterwards the RNA is precipitated and dissolved in selection buffer.

7.3.1.5 Selection of Aptamers

Aptamers can be selected in various ways. The most frequently used approaches are affinity chromatography [21] and modified cellulose filtration [4, 22]. The choice of method depends on the properties of the target (for example, its capability to be immobilized on a matrix or to be bound to modified cellulose filters) and the aim of selection. If the desired aptamers should, for example, bind molecules on the surface of intact cells, the selection scheme should employ these cells adhering to the surfaces of tissue culture flasks [23].

Other applicable selection methods include immunoprecipitation [24] and gel-shift assays [25]. Generally, a method must be found that enables the separation of unbound from bound RNA.

Renaturation of RNA

For renaturation, the purified RNA is dissolved in selection buffer after denaturing PAGE and precipitation (Section 7.3.1.4). Then this solution is heated to 70 °C for 5 min and subsequently cooled to the selection temperature within approximately 15 min. Alternatively, the RNA can be renatured, for example, by heating to other temperatures (65–95 °C) and immediately incubating it on ice. Mg^{2+} can be added either before or after heating. Because RNA can be degraded if heated in the presence of Mg^{2+}, the heating should not be too excessive if the buffer contains Mg^{2+} ions.

7.3.1.5.1 Affinity Chromatography

To perform affinity chromatography, the target molecules first have to be immobilized on the appropriate matrix.

Target Immobilization

Depending on the functional groups present in the target, a matrix must be chosen to which the target can be coupled in the most suitable way. Proteins are conveniently coupled by their NH_2, COOH or SH groups. Corresponding matrices, for example, are NHS-activated Sepharose 4 Fast Flow, EAH Sepharose 4B, or thiopropyl Sepharose 6B (Amersham Bioscience, Freiburg, Germany).

In a SELEX experiment, it should be possible to adjust the concentration of the immobilized target. It is also essential to know the stability of the bond between target and matrix to estimate the maximal storage period for the affinity material.

If sufficient target is available, new matrix can be used in every selection round. Otherwise, the matrix must be treated so that no denaturation or degradation of the target molecules occurs. Correspondingly, elution of bound RNA and regeneration of the matrix have to be carried out.

Preselection

To avoid the enrichment of RNA molecules that bind to pure matrix, preselection with matrix carrying no target should be performed. The renatured RNA is allowed to pass through a corresponding column or is simply incubated together with the matrix in batch. The preselection should differ from the main selection only in the absence of the immobilized target.

If an aptamer is desired that binds one molecule with high specificity, but should not bind a second, very similar molecule, this can be achieved by using this second molecule immobilized to the matrix during the preselection [10]. This process is referred to as counterselection.

By washing the preselection matrix with a few column volumes of selection buffer, RNA that is not bound to the matrix or only weakly bound is eluted. The preselected RNA pool is used in the following main selection.

With radioactively labeled RNA, the amount of RNA bound to the matrix can easily be detected by a beta counter. This facilitates the detection of a possible enrichment of matrix-binding RNA molecules.

Main Selection

The preselected RNA pool is either allowed to flow through a column with immobilized target (affinity chromatography) or is incubated with the matrix carrying the target (batch procedure).

In the latter approach, incubation times of about 5–60 min are common. To remove unbound RNA, the mixture is transferred into a column or funnel with an appropriate frit, unless the incubation was done in a vessel suitable for physical separation. Unbound RNA is eluted by gravity flow. Then the matrix is washed with several column volumes of selection buffer to enrich the strongest-binding molecules. In the first rounds of selection, the washing does not need to be performed as extensively as in the later rounds because usually less than 1% of the input RNA remains bound to the matrix after a few column volumes of washing.

The bound RNA may then be eluted by unspecific denaturation with 7 M urea, 0.5 mM EDTA, 2% SDS, or heat. Another possibility is specific recovery of the bound RNA by affinity elution, in which the matrix is incubated with selection buffer containing free target molecules in excess relative to the immobilized target, so that target-binding RNA molecules are eluted.

Enrichment of aptamers with a low k_{off} value is achieved by washing the matrix with a buffer containing free target prior to elution [26]. Aptamers with high k_{off} values dissociate quickly from the immobilized target and bind again preferably to free target molecules, which must be applied in excess relative to the immobilized target. Aptamers with low k_{off} rates do not dissociate as quickly from the immobilized target and thus are enriched.

If the eluted RNA is radioactively labeled, its amount can be determined, and the enrichment of aptamers can easily be followed. The SELEX experiment can be stopped if more than 30% of the RNA remains bound to the matrix despite intensive washing.

The Target Concentration

As mentioned in Section 7.3, the concentration of immobilized target plays an important role in a SELEX experiment. At the beginning of a selection, the target concentration is relatively high to make sure that a lot of binding RNA molecules will survive the actual round. However, in order to enrich the best-binding molecules the target concentration needs to be decreased as the selection progresses.

7.3.1.5.2 Modified Cellulose Filter Binding

The modified cellulose filter binding separation method is based on the ability of proteins to bind to nitrocellulose and cellulose acetate [27]. This technique is suitable for protein targets or targets that are linked to a protein, for example, biotin-labeled molecules that can form a stable complex with streptavidin, which itself is retained on modified cellulose [28].

Similar to the chromatographic selection step, the enrichment of undesired RNA molecules, which may bind to the filter, streptavidin, or other components of the reaction setup, must be prevented by preselection. For preselection, the RNA pool is passed through a cellulose filter. Unbound RNA is washed from the filter in a small volume of selection buffer.

The preselected pool is then incubated with the target molecule for 5–60 min, that is, similar to the chromatographic selection. Afterwards this mixture is vacuum-filtered through modified cellulose. The filter is then washed with selection buffer, and the bound RNA is eluted with a denaturing elution buffer, for example, 7 M urea, 0.5 mM EDTA, and by heating to 70 °C.

With this technique, a selection of aptamers with low k_{off} rates can be achieved by diluting the mixture of target and RNA before filtration [18]. Aptamers with high k_{off} rates dissociate from the complex with the target and do not reassociate with target molecules because of the decreased target concentration. Complexes of targets and aptamers with low k_{off} rates do not dissociate rapidly, leading to enrichment in 'low k_{off}-rate aptamers'.

7.3.1.6 Reverse Transcription and PCR

The selected RNA must be reverse-transcribed into cDNA for further amplification by PCR. Therefore, the RNA is precipitated and subjected to a reverse transcription reaction according to the protocol of the reverse transcriptase supplier. If the RNA concentration is very low (as in the first selection rounds) a coprecipitant like glycogen has to be added to recover the RNA quantitatively.

DNA-Template Amplification after Reverse Transcription

The PCR amplification after RT is started by adding an aliquot of the RT reaction to the prepared PCR mixture. If the amount of RNA used in the RT is known, it is possible to estimate the required number of PCR cycles. Otherwise, the progress of the

PCR reaction has to be monitored by agarose gel electrophoreses. After completing the PCR, the dsDNA product can simply be purified by ethanol precipitation, and dissolved in ddH_2O.

It is always recommended to examine whether the selected RNA or any used RT–PCR solution or equipment contains DNA contaminants, which would be preferably amplified in the PCR. Therefore, it is necessary to run controls without RT and without template.

The control without RT contains a small portion of the selected RNA and is treated like the rest of the RNA, by adding the RT reaction mixture but no reverse transcriptase. If a PCR product is observed in this control without RT, the product in the PCR reaction does not result solely from selected RNA molecules, but from pool DNA contaminating the RNA fraction (see Section 7.3.1.4).

The control reaction without template is carried out by not adding an aliquot of the selected RNA to the RT reaction. A PCR product in the control without template would suggest contamination of RT or PCR component or other laboratory equipment (for example, pipettes) with pool DNA.

Any DNA contamination must be removed because it is nearly impossible to enrich aptamers under these conditions.

7.3.2 Selection of 2′-Modified RNA Aptamers

The insufficient stability of RNA, mentioned above, limits the use of RNA aptamers. However, some established techniques yield nuclease-resistant aptamers. One alternative is called the Spiegelmer approach [29], another method involves incorporation of 2′-methoxy purine nucleotides [30].

The most common method employs the substitution of all pyrimidine nucleotides by their 2′-NH_2- or 2′-F-analogs [31]. These confer increased stability to the RNA, because most ribonucleases need the 2′-OH for cleaving RNA. Substitution of only the pyrimidine nucleotides increases the half-life of RNA from seconds to days [28]. Such 2′-modified RNAs are suitable for the detection of analytes in biological samples like blood, serum, or urine.

The selection of 2′-modified (pyrimidine) RNA aptamers is done exactly as described for RNA aptamers (see Section 7.3.1), except that 2′-modified UTP and CTP are used in transcription. The resulting modified RNA is also compatible with standard reverse transcription protocols.

In-vitro Synthesis of 2′-Modified RNA

The yield of *in vitro* synthesis of RNA from 2′-modified nucleotides is decreased 2–100-fold, compared to the reaction with unmodified nucleotides [32]. Various transcription buffers are reported to enhance the transcription by T7 RNA polymerase using 2′-modified pyrimidine nucleotides [28, 33]. These buffers mostly contain higher spermidine concentrations and additives like PEG or Triton. Additionally, the concentration of T7 RNA polymerase should be increased to improve

the yield of RNA. Also, variants of T7 RNA polymerase have been reported that incorporate modified nucleotides with higher efficiency than the wild type [34,35]. Furthermore, the number of pyrimidines in the first 12 base positions of the RNA transcript should be very low, to increase the transcript yield.

7.3.3 Selection of ssDNA Aptamers

As mentioned above, aptamers can also be made of DNA. Their selection (Figure 7.3) differs slightly from an RNA aptamer selection (Figure 7.2). Instead of the *in vitro* transcription of DNA into RNA, ssDNA is prepared as described in Section 7.3.3.1. Of course, inclusion of an RNA polymerase promoter in the template design, as well as the reverse transcription step, are not necessary. All other steps are the same as those for RNA aptamer selection (see Section 7.3.1).

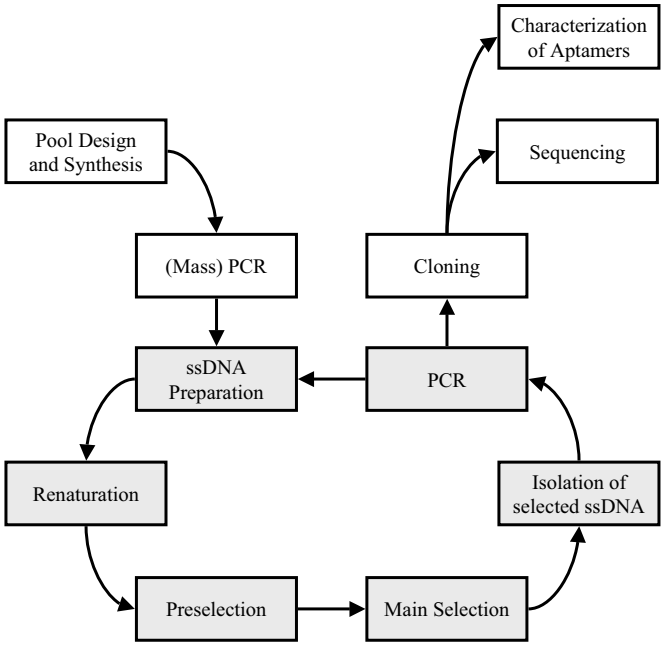


Fig. 7.3. *In vitro* selection of an ssDNA aptamer. Steps of one cycle are given in gray.

7.3.3.1 Generation of ssDNA

Generation of ssDNA is necessary for ensuring that aptamers can fold into unique tertiary structures. In the presence of a complementary strand, a dsDNA helix would form, which hinders selection. The starting material is an ssDNA pool, which is am-

plified by PCR to yield a dsDNA pool (Figure 7.2). The ssDNA aptamer is created by separating the complementary strands of the dsDNA pool. Physical separation of the two strands can be achieved by the addition of biotin to one strand and preparative PAGE or by capturing the biotinylated strand with immobilized streptavidin [36]. The biotin is introduced into the DNA by using a 5′-biotinylated primer. The unlabelled strand usually serves as the ssDNA pool for selecting aptamers.

7.3.3.2 Selection of Aptamers

After the ssDNA pool has been established, it must be de- and renatured as described for RNA aptamers (see Section 7.3.1.5). Then preselection and selection are done, and the selected ssDNA is used in a PCR to create the new dsDNA pool of the next selection round.

7.3.4 Cloning and Sequencing

When the selection is finished, individual aptamers are identified by cloning in *Escherichia coli*, followed by sequence analysis. Therefore, plasmids are prepared from single colonies to obtain 'monoclonal' templates for the sequencing and preparative-scale production of the aptamers. For this purpose, we use the TOPO TA Cloning® Kit (Invitrogen, Karlsruhe, Germany) for fast and convenient cloning. Alternatively, the aptamer pool may be cloned after digestion with appropriate restriction endonucleases and ligation into a plasmid that contains the corresponding recognition sites. However, this cloning strategy must be planned when designing the starting pool (see Section 7.3.1.1). The base sequences are determined by any of the established sequencing methods [14].

7.3.5 Characterization of Aptamers

The methods described below aim at determining the binding properties of the selected aptamers (K_D, k_{on}, k_{off}). Other methods of characterizing aptamers further include, for example, sequence alignment to identify common sequence motifs (in different aptamers resulting from the same selection, from different selections, or from other aptamers and natural RNA sequences). Secondary structures can be determined by computational prediction [37–39] (see also Chapter 12), as well as by structure mapping [40, 41]. Similar to footprinting assays, these analyses may help to identify binding motifs and to create minimal aptamers.

7.3.5.1 Surface Plasmon Resonance Spectroscopy

With surface plasmon resonance spectroscopy, k_{on} and k_{off} rates of the aptamer-target complex are determined. These in turn, allow calculation of the K_D value [42]. For surface plasmon resonance spectroscopy, one of the two binding partners

is immobilized on the surface of a sensor chip. If the target is the immobilized component, several aptamers can be measured with just one chip. If it is impossible to immobilize the target, every single aptamer must be immobilized separately to investigate its interaction with the target. Thus, this approach does not allow a throughput as high as the first approach.

The main difficulty with this technology is the possible interaction of target and/or aptamer with the chip surface. The nature of this difficulty may be high unspecific interaction or high repulsion. For this reason, the most suitable sensor chip surface and optimal binding conditions must be determined for every target/aptamer combination.

7.3.5.2 Modified Cellulose Filter Binding Assay

The modified cellulose filter binding assay is based on the tight binding of proteins to this kind of filter material. When a protein-nucleic acid mixture is filtered, proteins are retained on the filter while nucleic acids are washed through. However, nucleic acids are also retained on the filter if they are bound to proteins. Thus, free and protein-bound nucleic acids can be separated [43].

By incubating very small amounts of radioactively labeled nucleic acids (at least 100 times less than the amount of protein) with increasing concentrations of a target protein, the K_D value can be calculated after quantification of the free nucleic acids (that is, those that are not bound to the filter) and the nucleic acid-target complexes that are retained on the filter.

7.3.5.3 Fluorescence Correlation Spectroscopy

The relatively new method of fluorescence correlation spectroscopy (FCS) is based on the fact that molecules with different molecular weights (usually) exhibit different diffusion times in solution. Thus, small molecules diffuse faster than larger ones. To determine K_D values, one component must be labeled with a fluorescent dye. Due to the different molecular weights of the uncomplexed, labeled component and the complex, the diffusion times of the free and complexed molecule differ. This fact allows determining the distribution of free and complexed molecules in the solution. After measuring the distribution in different mixtures with varying ligand concentrations, the K_D value can be calculated [44].

If the target is fluorescently labeled, different aptamers can be measured by using the same labeled ligand. If it is impossible to label the target, every single aptamer must be labeled for each measurement.

Some problems with FCS technology concern the fluorescent label. Data collection is possible only if the label does not interfere with the structure of the labeled molecule or the binding of the target to the aptamer. Also, the label should be site-specific, because molecules that are labeled at different positions may have different binding properties. Another problem concerns the molecular weight difference between the free and complexed molecule, which should be sufficient for

a significant variation in the diffusion times. Therefore, this method is well suited if the target molecule can be fluorescently labeled and if its size is significantly smaller than the corresponding complex.

7.3.5.4 Gel-shift Assay

In the gel-shift assay which is very similar to the modified cellulose filter binding assay (see Section 7.3.5.2), free and complexed aptamers are separated by native PAGE [42]. Small amounts of radioactively labeled aptamers (100 times less than the target molecules) are incubated with increasing target concentrations. After native PAGE the amounts of free and complexed aptamers (= shifted band) are determined by autoradiography. Based on the law of mass action, the K_D value can be determined.

The problem with gel-shift assays concerns the fact that the interactions under investigation must be very strong (low k_{off} rates). Weakly interacting binding partners dissociate too quickly during electrophoresis and prevent determination of the K_D value.

7.3.6 Example: Isolation of Moenomycin A-specific Aptamers

In this section, we utilize the selection of moenomycin A-specific 2′-modified RNA aptamers by affinity chromatography [28] to give a more detailed overview of a SELEX experiment.

7.3.6.1 Immobilization of Target Molecules

For chromatographic selection, moenomycin A-derivatized matrix was produced by incubating prepared thiol sepharose (described in [28]) with moenomycin-S-S-pyridone for 3 h at 4 °C by gently swirling in immobilization buffer (Figure 7.4).

The released 2-thiopyridone was quantified by UV spectroscopy (343 nm) to determine the amount of immobilized moenomycin A.

The moenomycin-sepharose was stored in selection buffer at 4 °C and could be used for about 2 weeks.

7.3.6.2 PCR

The reaction was carried out in five 100 µL mixtures containing 1× PCR buffer, 200 µM of each dNTP, 1 µM primer A, 1 µM primer B, 2 U DNA polymerase, and 10 µL of the RT mixture. In the first SELEX round, the ssDNA pool was used as the template.

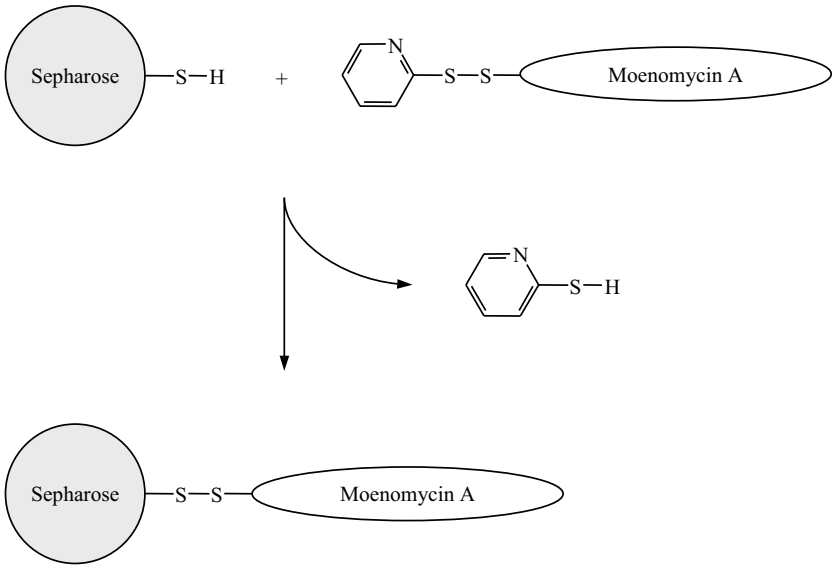


Fig. 7.4. Immobilization of moenomycin A on thiol Sepharose.

PCR profile: 1 min 94 °C, 1 min 55 °C, 1 min 72 °C.

The dsDNA product was ethanol-precipitated and dissolved in ddH_2O.

7.3.6.3 *In vitro* Transcription

Transcription mixture (100 µL) containing 1× transcription buffer, 1 mM 2′-NH_2-CTP, 1 mM 2′-NH_2-UTP, 3 mM ATP, 3 mM GTP, 10 µCi [α-^{32}P]-GTP, 500 U T7 RNA polymerase, and 200 pmol dsDNA was prepared and incubated for 4 h at 37 °C.

In the first SELEX round 6 transcriptions were performed to yield 7.4 nmol RNA.

7.3.6.4 RNA Purification

DNase I Digestion

After the *in vitro* transcription was finished, 5 U DNase I was added to the mixture, which was then incubated for 10 min at 37 °C.

RNA Purification

The RNA yielded by *in vitro* transcription was purified by denaturing PAGE (8% polyacrylamide (acrylamide/bisacrylamide 19:1), 7 M urea).

Full-length RNA was detected by UV-shadowing, eluted from the gel by the 'crush and soak' method [14], and ethanol-precipitated.

7.3.6.5 Selection of Aptamers

De- and Renaturation

The RNA pellet was dissolved in the selection buffer, heated at 70 °C for 10 min. Then, Mg^{2+} was added and the RNA was renatured by allowing the mixture to cool to room temperature within 30 min.

Preselection

Thiol Sepharose 4B (200 µL) was incubated with the RNA by gently mixing at room temperature for one hour. This mixture was then transferred into a chromatography column, and the unbound RNA was eluted by gravity flow.

Main Selection

Thiol Sepharose 4B derivatized with moenomycin A was incubated with the preselected RNA pool for 1 h at room temperature while being mixed gently. This mixture was then transferred to an empty chromatography column. Unbound RNA was removed from the matrix by washing with selection buffer. Bound RNA was eluted by cleaving the moenomycin A from the Sepharose by washing the column 5 times with 200 µL selection buffer containing 200 mM DTT.

Details and the results of each SELEX round are shown in Table 7.1.

Table 7.1. Selection of moenomycin A-specific aptamers.

SELEX round	1	2	3	4	5	6	7	8	9	10	11
RNA presel. column [%]*	0.6	1.8	1.9	1.9	2.2	1.4	0.5	1.9	0.8	1.3	1.8
Moenomycin sepharose [µL]	400	200	100	100	100	50	50	50	50	50	25
Moenomycin [µM]	930	930	930	930	930	930	300	300	130	70	70
Wash step [column volumes]	4	4	4	7	1050	100	100	100	10	10	30
RNA selected [%]	0.2	0.6	1.0	0.2	8.6	18.2	2.3	10.0	5.6	3.4	26.5

* RNA bound to the preselection column (%)

7.3.6.6 Reverse Transcription

The selected RNA was ethanol-precipitated in the presence of glycogen (100 µg L^{-1}). The RNA pellet was dissolved in 32.5 µL ddH_2O and 0.5 µL primer B (200 µM) was added.

This mixture was heated for 1 min to 95 °C and was then immediately incubated on ice. Then 10 µL 5× RT buffer, 2 µL DTT (100 mM), 4 µL dNTP-mix (4 mM of each dNTP), and 1 µL Super Script^{tm} II reverse transcriptase (200 U µL^{-1}) was added. Reverse transcription was performed for 1 h at 42 °C. This mixture was subsequently used in the PCR (see above). After round 11, several aptamers were isolated and identified by cloning and sequencing [28].

7.4 Troubleshooting

Specific steps that can go wrong have already been discussed in the related sections. Here, some more general problems are mentioned.

RNA Stability

Care must be taken to prevent rapid degradation of RNA samples, because ribonucleases (RNases) are nearly ubiquitous. Thus, it is strongly recommended to use certified RNase-free reagents, enzymes, and equipment (for example, tubes) as well as working 'clean'. Further arrangements include the addition of RNase inhibitors to reaction mixtures, using diethylpyrocarbonate (DEPC) treatment to remove RNases from water, buffers, and other solutions, and the indication of special 'RNA zones' in the laboratory that are exclusively assigned to RNA experiments. Potential RNase contamination can be detected by incubating intact RNA with all suspicious solutions and components for one hour and examining for degraded RNA by gel electrophoresis.

RNA is also nonenzymatically degraded by hydroxyl ions and, at high temperatures, by divalent cations. Hence, solutions that contain RNA should not exceed a pH value of 9 and should not be heated for too long periods of time if divalent cations are present.

Enrichment of Low-affinity Aptamers

To prevent the selection of aptamers with low affinity to the target molecule, the target concentration has to be decreased and the washing procedure has to be done with increased stringency. This enhances the discrimination between aptamers with high affinity and aptamers with lower affinity [6].

Enrichment of Matrix-binding Aptamers

Despite a preselection procedure, aptamers may be selected that bind to the bare selection matrix, the cellulose filter, or any other component of the selection setup, because the possible binding sites are presented in a large number (compared to

the selection target). If a selection procedure leads to an enrichment of matrix-binding aptamers, the preselection procedure should be done more stringently or the separation method should be changed during a SELEX experiment (for example, from filter-binding assay to gel-shift assay). On the other hand, a too stringent main selection can also facilitate the enrichment of matrix-binding aptamers, if the number of selected target-binding molecules is too small [6].

7.5 Major Applications

Aptamers can be employed in many different analytical and, possibly, therapeutic applications [45]. They can replace (monoclonal) antibodies in nearly all instances in ELISAs [46] and immunohistological applications [36]. The relative instability of RNA can be circumvented by various strategies (see Section 7.3.2). One possible application is the use of intramers [47]. This name is derived from 'intracellularly produced aptamers'. These aptamers are selected for binding an intracellular molecule like a signaling receptor, and they are able to inhibit the biological function of their target molecules while they are produced *in vivo*. Aptamers can also be used to detect analytes in various samples by measuring the change in mass on a sensor surface (caused by binding of the target to the immobilized aptamer [48]) or the change in fluorescence anisotropy [49, 50] or intensity [13, 51] (caused by ligand binding). With reporter ribozymes [52], a change in fluorescence intensity is monitored.

Also, high-throughput screening of small ligands for proteins [53] or other biomacromolecules of therapeutic interest can be done with the aid of aptamers. Because small molecules replace an aptamer in its complex with the macromolecule, this substitution can be monitored by a change of fluorescence anisotropy of labeled aptamer or by a change in the activity of an aptazyme.

Acknowledgments

We thank Daniela Otto and Rico Czaja for critically reading the manuscript and the German Fonds der Chemischen Industrie for financial support.

References

1. Ellington, A.D. and Szostak, J.W. (**1990**) In vitro selection of RNA molecules that bind specific ligands, *Nature 346*, 818–822.
2. Seelig, B. and Jaschke, A. (**1999**) A small catalytic RNA motif with Diels-Alderase activity, *Chem Biol. 6*, 167–176.

3. Li, Y. and Breaker, R.R. (**2001**) In vitro selection of kinase and ligase deoxyribozymes, *Methods 23*, 179–190.
4. Tuerk, C. and Gold, L. (**1990**) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase, *Science 249*, 505–510.
5. Bacher, J.M. and Ellington, A.D. (**1998**) Nucleic acid selection as a tool for drug discovery, *Drug Discovery Today 3,* 265–273.
6. Ellington, A.D. and Conrad, R. (**1995**) Aptamers as potential nucleic acid pharmaceuticals, *Biotechnol Annu Rev. 1*, 185–214.
7. Geiger, A., Burgstaller, P., von der Eltz, H., Roeder, A. and Famulok, M. (**1996**) RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity, *Nucleic Acids Res. 24*, 1029–1036.
8. Bridonneau, P., Chang, Y.F., O'Connell, D., Gill, S.C., Snyder, D.W., Johnson, L., Goodson, T. Jr., Herron, D.K. and Parma, D.H. (**1998**) High-affinity aptamers selectively inhibit human nonpancreatic secretory phospholipase A2 (hnps-PLA2), *J. Med. Chem. 41*, 778–786.
9. Mayer, G., Blind, M., Nagel, W., Bohm, T., Knorr, T., Jackson, C.L., Kolanus, W. and Famulok, M. (**2001**) Controlling small guanine-nucleotide-exchange factor function through cytoplasmic RNA intramers, *Proc. Natl. Acad. Sci. USA 98*, 4961–4965.
10. Jenison, R.D., Gill, S.C., Pardi, A. and Polisky, B. (**1994**) High-resolution molecular discrimination by RNA, *Science 263*, 1425–1429.
11. Cox, J.C. and Ellington, A.D. (**2001**) Automated selection of anti-protein aptamers, *Bioorg. Med. Chem. 9*, 2525–2531.
12. Rogers, J. and Joyce, G.F. (**1999**) A ribozyme that lacks cytidine, *Nature 402*, 323–325.
13. Jhaveri, S., Rajendran, M. and Ellington, A.D. (**2000**) In vitro selection of signaling aptamers, *Nat. Biotechnol. 18*, 1293–1297.
14. Sambrook, J. and Russell, D.W. (**2001**) *Molecular Cloning: A Laboratory Manual*, Third edn, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
15. Shtatland, T., Gill, S.C., Javornik, B.E., Johansson, H.E., Singer, B.S., Uhlenbeck, O.C., Zichi, D.A. and Gold, L. (**2000**) Interactions of *Escherichia coli* RNA with bacteriophage MS2 coat protein: genomic SELEX, *Nucleic Acids Res. 28*, E93.
16. Bittker, J.A., Le, B.V. and Liu, D.R. (**2002**) Nucleic acid evolution and minimization by nonhomologous random recombination, *Nat. Biotechnol. 20*, 1024–1029.
17. Coleman, T.M. and Huang, F. (**2002**) RNA-catalyzed thioester synthesis, *Chem. Biol. 9*, 1227–1236.
18. Marshall, K.A. and Ellington, A.D. (**2000**) In vitro selection of RNA aptamers, *Methods Enzymol. 318*, 193–214.
19. Milligan, J.F. and Uhlenbeck, O.C. (**1989**) Synthesis of small RNAs using T7 RNA polymerase, *Methods Enzymol. 180*, 51–62.
20. Hassur, S.M. and Whitlock, H.W., Jr. (**1974**) UV shadowing: a new and convenient method for the location of ultraviolet-absorbing species in polyacrylamide gels, *Anal. Biochem. 59*, 162–164.
21. Proske, D., Gilch, S., Wopfner, F., Schatzl, H.M., Winnacker, E.L. and Famulok, M. (**2002**) Prion-protein-specific aptamer reduces PrPSc formation, *Chembiochem. 3*, 717–725.
22. White, R.R., Shan, S., Rusconi, C.P., Shetty, G., Dewhirst, M.W., Kontos, C.D. and Sullenger, B.A. (**2003**) Inhibition of rat corneal angiogenesis by a nuclease-resistant RNA aptamer specific for angiopoietin-2, *Proc. Natl. Acad. Sci. USA 100,* 5028–5033.
23. Hicke, B.J., Marion, C., Chang, Y.F., Gould, T., Lynott, C.K., Parma, D., Schmidt, P.G. and Warren, S. (**2001**) Tenascin-C aptamers are generated using tumor cells and purified protein, *J. Biol. Chem. 276*, 48644–48654.

24. Tsai, D.E., Harper, D.S. and Keene, J.D. (**1991**) U1-snRNP-A protein selects a ten nucleotide consensus sequence from a degenerate RNA pool presented in various structural contexts, *Nucleic Acids Res. 19*, 4931–4936.
25. Blackwell, T.K. and Weintraub, H. (**1990**) Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection, *Science 250*, 1104–1110.
26. Davis, J.H. and Szostak, J.W. (**2002**) Isolation of high-affinity GTP aptamers from partially structured RNA libraries, *Proc. Natl. Acad. Sci. USA 99*, 11616–11621.
27. Tovey, E.R. and Baldo, B.A. (**1989**) Protein binding to nitrocellulose, nylon and PVDF membranes in immunoassays and electroblotting, *J. Biochem. Biophys. Methods 19,* 169–183.
28. Schürer, H., Stembera, K., Knoll, D., Mayer, G., Blind, M., Forster, H.H., Famulok, M., Welzel, P. and Hahn, U. (**2001**) Aptamers that bind to the antibiotic moenomycin A, *Bioorg. Med. Chem. 9*, 2557–2563.
29. Klussmann, S., Nolte, A., Bald, R., Erdmann, V.A. and Furste, J.P. (**1996**) Mirror-image RNA that binds D-adenosine, *Nat. Biotechnol. 14*, 1112–1115.
30. Green, L.S., Jellinek, D., Bell, C., Beebe, L.A., Feistner, B.D., Gill, S.C., Jucker, F.M. and Janjic, N. (**1995**) Nuclease-resistant nucleic acid ligands to vascular permeability factor/vascular endothelial growth factor, *Chem. Biol. 2*, 683–695.
31. Pagratis, N.C., Bell, C., Chang, Y.F., Jennings, S., Fitzwater, T., Jellinek, D. and Dang, C. (**1997**) Potent 2′-amino-, and 2′-fluoro-2′-deoxyribonucleotide RNA inhibitors of keratinocyte growth factor, *Nat. Biotechnol. 15*, 68–73.
32. Aurup, H., Williams, D.M. and Eckstein, F. (**1992**) 2′-Fluoro- and 2′-amino-2′-deoxynucleoside 5'-triphosphates as substrates for T7 RNA polymerase, *Biochemistry 31*, 9636–9641.
33. Jellinek, D., Green, L.S., Bell, C., Lynott, C.K., Gill, N., Vargeese, C., Kirschenheuter, G., McGee, D.P., Abesinghe, P., Pieken, W. A. and et al. (**1995**) Potent 2′-amino-2′-deoxypyrimidine RNA inhibitors of basic fibroblast growth factor, *Biochemistry 34*, 11363–11372.
34. Sousa, R. (**2000**) Use of T7 RNA polymerase and its mutants for incorporation of nucleoside analogs into RNA, *Methods Enzymol. 317*, 65–74.
35. Padilla, R. and Sousa, R. (**2002**) A Y639F/H784A T7 RNA polymerase double mutant displays superior properties for synthesizing RNAs with non-canonical NTPs, *Nucleic Acids Res. 30*, e138.
36. Blank, M., Weinschenk, T., Priemer, M. and Schluesener, H. (**2001**) Systematic evolution of a DNA aptamer binding to rat brain tumor microvessels. selective targeting of endothelial regulatory protein pigpen, *J. Biol. Chem. 276*, 16464–16468.
37. Zuker, M. (**1989**) On finding all suboptimal foldings of an RNA molecule, *Science 244*, 48–52.
38. Zuker, M. (**1989**) Computer prediction of RNA structure, *Methods Enzymol. 180*, 262–288.
39. Zuker, M. (**2003**) Mfold web server for nucleic acid folding and hybridization prediction, *Nucleic Acids Res. 31*, 3406–3415.
40. Stern, S., Weiser, B. and Noller, H.F. (**1988**) Model for the three-dimensional folding of 16 S ribosomal RNA, *J. Mol. Biol. 204*, 447–481.
41. McGregor, A., Murray, J.B., Adams, C.J., Stockley, P.G. and Connolly, B. A. (**1999**) Secondary structure mapping of an RNA ligand that has high affinity for the MetJ repressor protein and interference modification analysis of the protein-RNA complex, *J. Biol. Chem. 274*, 2255–2562.

42. Kim, S.J., Kim, M.Y., Lee, J.H., You, J.C. and Jeong, S. (**2002**) Selection and stabilization of the RNA aptamers against the human immunodeficiency virus type-1 nucleocapsid protein, *Biochem. Biophys. Res. Commun. 291*, 925–931.
43. Wong, I. and Lohman, T.M. (**1993**) A double-filter method for nitrocellulose-filter binding: application to protein-nucleic acid interactions, *Proc. Natl. Acad. Sci. USA 90,* 5428–5432.
44. Schürer, H., Buchynskyy, A., Korn, K., Famulok, M., Welzei, P. and Hahn, U. (**2001**) Fluorescence correlation spectroscopy as a new method for the investigation of aptamer/target interactions, *Biol. Chem. 382*, 479–481.
45. Famulok, M. and Mayer, G. (**1999**) Aptamers as tools in molecular biology and immunology, *Curr. Top. Microbiol. Immunol. 243*, 123–136.
46. O'Sullivan, C.K. (**2002**) Aptasensors: the future of biosensing? *Anal. Bioanal. Chem. 372*, 44–48.
47. Famulok, M., Blind, M. and Mayer, G. (**2001**) Intramers as promising new tools in functional proteomics, *Chem. Biol. 8*, 931–939.
48. Liss, M., Petersen, B., Wolf, H. and Prohaska, E. (**2002**) An aptamer-based quartz crystal protein biosensor, *Anal. Chem. 74*, 4488–4495.
49. Potyrailo, R.A., Conrad, R.C., Ellington, A.D. and Hieftje, G.M. (**1998**) Adapting selected nucleic acid ligands (aptamers) to biosensors, *Anal. Chem. 70*, 3419–3425.
50. McCauley, T.G., Hamaguchi, N. and Stanton, M. (**2003**) Aptamer-based biosensor arrays for detection and quantification of biological macromolecules, *Anal. Biochem. 319*, 244–20.
51. Stojanovic, M.N., de Prada, P. and Landry, D.W. (**2001**) Aptamer-based folding fluorescent sensor for cocaine, *J. Am. Chem. Soc. 123*, 4928–4931.
52. Hartig, J.S., Najafi-Shoushtari, S.H., Grune, I., Yan, A., Ellington, A.D. and Famulok, M. (**2002**) Protein-dependent ribozymes report molecular interactions in real time, *Nat. Biotechnol. 20*, 717–722.
53. Burgstaller, P., Girod, A. and Blind, M. (**2002**) Aptamers as tools for target prioritization and lead identification, *Drug Discovery Today 7*, 1221–1228.

8 Methods for Selecting Catalytic Nucleic Acids

Benjamin L. Holley, and Bruce E. Eaton

8.1 Introduction

The enormous potential of nucleic acids (RNA and DNA) as catalysts is just beginning to be realized. With *in vitro* selection and amplification [1, 2] techniques, new nucleic acid catalysts have been discovered for a variety of reactions. Among biopolymers, nucleic acids are unique, being both directly enzymatically replicable and able to form stable tertiary structures containing catalytic active sites. However, in natural biological systems the vast majority of chemistry is catalyzed by proteins. In contrast to nucleic acids, proteins have diverse functional groups as side chains that can provide intricate interactions with substrates and cofactors [3]. For nucleic acids this lack of functional group diversity has been overcome by incorporation of modified nucleotides with a variety of functional groups, which extend even beyond the diversity present in proteins [4–6]. This equips nucleic acids with an expanded potential for catalysis. Since the underlying catalytic platform also carries sequence information that can be used for replication, catalysts with specific properties can be selected for and amplified from large pools of random RNA sequence [7, 8].

Modified and unmodified RNA and DNA catalysts have been discovered through *in vitro* selection techniques. RNA was the first type of nucleic acid found to have catalytic properties. The first 'ribozyme' was an unmodified nuclear RNA found to catalyze site-specific phosphodiester transfer reactions, resulting in self-cleavage and ligation [9]. More recently, RNA has been found to catalyze many other types of reactions, such as urea synthesis [10], Diels-Alder cycloaddition [7, 11], amidation [12], and esterification [13, 14]. Although not naturally occurring, DNA catalysts have also been produced that can cleave the phosphodiester bond of oligoribonucleotides [15] and facilitate the metallation of porphyrins [16]. DNA and RNA have similar native functional groups and possibilities for base modifications; therefore, they have equal potential as a catalytic platform [17]. Because there are more examples to draw from when designing an *in vitro* RNA selection, RNA is the main focus of this chapter.

The process of selecting an RNA catalyst with a particular function begins with a large pool of random sequences. This pool can be acquired by producing random single-stranded DNA on a nucleic acid synthesizer or by *in vitro* mutagenesis of an existing nucleic acid library [18, 19]. Random sequence pools produced by chemical

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst
Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-30799-0

synthesis are potentially the most diverse ($\sim 10^{14}$ different sequences), but can be limited in size to approximately 200 bases, due to low overall coupling efficiencies during chemical synthesis. Mutagenic PCR can also be used to randomize nucleic acid sequences, but not to the same extent as chemical synthesis. Mutagenesis has the advantage of allowing one to fine-tune the degree of mutation introduced into a particular sequence or pool of sequences. Another way of introducing diversity into nucleic acid libraries is chemical modification of the bases constituting the oligonucleotide. Together, these sources of sequence complexity and functional group diversity can give rise to enormous pools from which to select potential catalysts.

From RNA pools of random sequence, active catalysts can be partitioned from the bulk of the pool based on their ability to accomplish a particular catalytic task. The substrates for many of these catalysts are templated [20], meaning that they contain a nucleic acid portion that can hydrogen-bond in predetermined Watson-Crick fashion with a complementary region of the catalyst, thereby facilitating the initial substrate-binding step of catalysis. Other substrates not templated have also been used for RNA catalysis. These examples can be free in solution or tethered to the catalyst itself through a flexible linker [7,21,22]. Partitioning of the active RNA catalysts away from the starting pool is most conveniently facilitated by having one of the substrates covalently attached to the catalyst. This can be accomplished by a variety of methods that are discussed later. By selectively capturing the product, the catalyst that produced it can be isolated and amplified by PCR (Figure 8.1). This type of process for generation of RNA catalysts has been the most widely studied. The many variations in modification of either the substrate or the RNA must be considered carefully when designing a catalyst *in vitro* selection experiment.

8.2 Materials and Equipment

Producing the Random Library

- Automated nucleic acid synthesizer
- Thermocycler
- Taq DNA polymerase (NEB)
- Pfu DNA polymerase (PfuUltra, Stratagene)
- 3′ and 5′ primers, HPLC purified (Operon)
- 6% native PAGE solution (40 mL)
- Electrophoresis apparatus (1.5 mm × 16 cm × 14 cm)
- Ethidium bromide stain (1 μg mL^{-1} H_2O)
- UV transilluminator

Purification of Libraries by Electrophoresis

- 40% acrylamide (19:1 bisacrylamide)

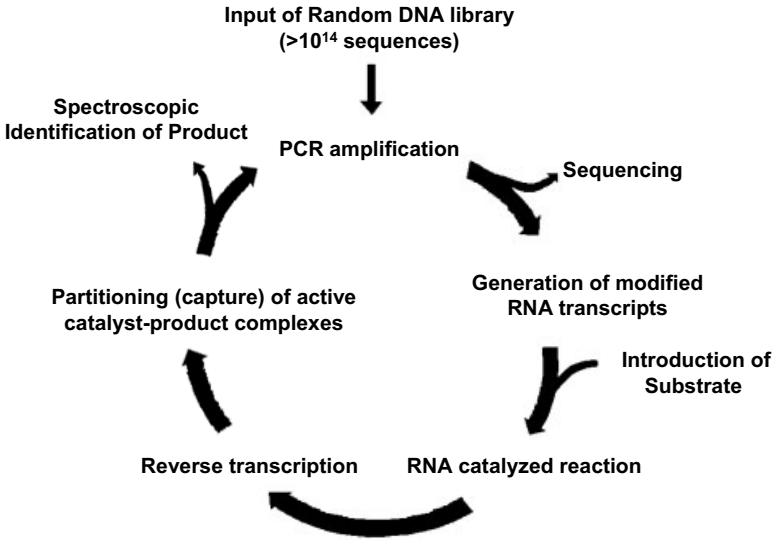


Fig. 8.1. Generalized selection cycle for *in vitro* evolution of an RNA catalyst. Random libraries are PCR-amplified, transcribed, modified with a tethered reactant, reacted with a second substrate in solution, and reverse-transcribed. Active RNA/cDNA library constructs are separated from inactive ones so that they can enter the next cycle of selection.

- 10× TBE (Tris-borate-EDTA):
  Tris-Cl (pH 8.3) 1.1 M
  boric acid 0.9 M
  EDTA 25 mM
- 2× native loading buffer (10 mL):
  glycerol 1 mL
  10× TBE 2 mL
  bromophenol blue 0.025% (w/v)
  xylene cyanol 0.025% (w/v)
  H_2O up to 10 mL
- 2× denaturing loading buffer (10 mL):
  formamide 9.5 mL
  EDTA (0.5 M) 200 µL
  bromophenol blue 0.025% (w/v)
  xylene cyanol 0.025% (w/v)
  H_2O up to 10 mL
- Electrophoresis apparatus (1.5 mm × 16 cm × 14 cm)
- Silica DNA purification column (QIAquick, Qiagen)

Transcription

- ATP, GTP, CTP, and UTP or modified UTP
- [α-^{32}P] ATP, 3000 Ci µM^{-1} (ICN)

- RNase inhibitor (RNasin, Promega)
- T7 RNA polymerase (Promega)
- Silica RNA purification column (RNeasy, Qiagen)

Ligation

- T4 DNA ligase (Promega)
- 20-mer DNA-bridging oligonucleotide (Operon)
- 10-mer DNA-PEG substrate provided by custom synthesis
- RNase inhibitor (RNasin, Promega)

Purification of Ligation Products

- Thermocycler
- 6% denaturing PAGE solution (40 mL)
- Electrophoresis apparatus (1.5 mm × 16 cm × 14 cm)
- Phosphorimager
- 10× TB (Tris-borate)

Tris-Cl (pH 8.3)	1.1 M
boric acid	0.9 M

- Electroelution apparatus
- 30 kDa MWCO filters (Millipore)
- Sterile scalpel

Model RNA-catalyzed Reaction

- Argon (for metal storage)
- Anhydrous DMSO
- BMCC-biotin (Pierce)
- 30 kDa MWCO filters (Millipore)

Reverse Transcription

- Reverse transcriptase (SuperScript II, Invitrogen)
- 3′ primer, HPLC purified (Operon)
- RNase inhibitor (RNasin, Promega)

Purification after Reverse Transcription

- Activated silica resin (StrataClean, Stratagene)
- 30 kDa MWCO filters (Millipore)
- 0.2 µm spin filter (Corning)

Partitioning

- Streptavidin, lyophilized (Scripps)
- Electrophoresis apparatus (1.5 mm × 16 cm × 14 cm)
- 6% denaturing PAGE solution (40 mL)

Bulk Sequencing

- Thermo Sequenase Cycle Sequencing Kit (Amersham)
- [γ-^{32}P] ATP, 5000 Ci μM^{-1} (ICN)
- T4 polynucleotide kinase (Promega)
- 3′ or 5′ primer, HPLC purified (Operon)

Cloning

- PCR-Script Amp Cloning Kit (Stratagene)

Additional Equipment

- Microcentrifuge (2 mL samples) capable of 12 000 $\times g$
- Benchtop centrifuge (30 mL samples) capable of 3000 $\times g$

8.3 Protocols

8.3.1 Generating the Starting Library

8.3.1.1 The Design

The ability to replicate nucleic acids directly is what makes an RNA catalyst *in vitro* selection possible. Utilizing the sequence information allows for enzymatic processes such as amplification, transcription, cleavage, and ligation to take place reproducibly. Unfortunately, the starting RNA pool cannot be composed completely of random sequences. Conserved regions of predetermined sequence must flank the random sequence region designed into the initial library. These conserved regions must contain sites for primer annealing and initiation of RNA transcription. In addition, these fixed regions can contain transcription promoters, sites for specific hybridization of substrates, and annealing sequences for the purpose of modification by ligation [7, 23]. The internal random region can vary in length (20 to >100 nucleotides). When designing the random region, it should be remembered that making this region larger decreases the probability of having all sequence possibilities represented completely [24]. It is important, however, to have sufficient length so that catalytically active tertiary structures can form. The practical production of libraries by solid-phase DNA synthesis puts a limit on the size of this region at about 200 bases.

The fixed regions provide the opportunity for significant engineering of the selection system. In addition to providing the initiation site for T7 RNA polymerase, the fixed region can be used for sequence-specific ligation (Section 8.3.3) of tethered substrate-DNA complexes (DNA-PEG substrate) to the RNA library by T4 DNA ligase. The presence of a bridging oligonucleotide that is complementary to both the substrate-DNA and the 5′ fixed region of the RNA pool facilitates this

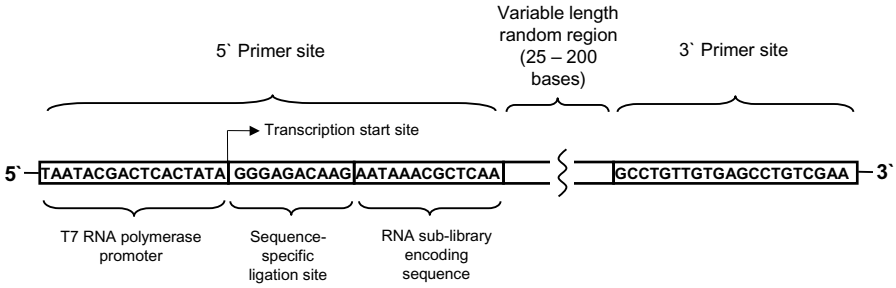


Fig. 8.2. Design of a synthetic DNA random library template for use as starting material for *in vitro* selection. Various sites are engineered into this construct to allow for PCR by Taq polymerase, transcription by T7 RNA polymerase, and ligation with T4 DNA ligase. This construct also includes a 100-nucleotide region of random sequence that will become the evolved catalytic region.

process (as discussed later). This allows for sublibraries of potential catalysts with sequence-specific tethered substrates to be produced and used as mixtures in an *in vitro* selection. The fate of various substrates and the catalysts that carried out their transformations can be tracked in large pools by quantitative PCR and recovered by selective amplification [25]. This greatly increases the various modes of potential RNA catalysis that can be explored in a single selection. Other fixed-region sequences that can be included are sites for restriction endonuclease digestion to aid in cloning or separation of products for analysis. Fixed regions may also be placed inside the random region as an internal frame marker. A generalized DNA construct containing most of the features mentioned above is shown in Figure 8.2.

8.3.1.2 Producing a Random Pool

Once designed, a RNA pool can be readily made by automated solid-phase DNA synthesis on an Applied Biosystems (ABI 394) nucleic acid synthesizer according to the manufacturers protocol [18]. There is a bias toward the coupling of some nucleobase phosphoramidites over others, so nonequimolar ratios are used. Synthesis of the random region can be accomplished by using a 3:3:2:2 ratio (A:C:G:T) of the four deoxyribonucleoside phosphoramidites at each coupling step. Depending on the size of the random region, yields can be low for synthesis. Elimination of the capping step during random region synthesis has been reported to improve yields [22]. After standard deprotection and desalting steps, the synthetic library can be purified by denaturing polyacrylamide gel electrophoresis and excision of the desired-length DNA followed by electroelution (Section 8.3.1.5). Gel-purified pools can be amplified by PCR (Section 8.3.1.3) to obtain the quantities needed for the selection. It should be noted that the quantity of random sequence space available for *in vitro* selection is limited by the amount of DNA isolated after PCR.

Random pools can also be generated by using various forms of mutagenic PCR (Section 8.3.1.4). The pools derived from these methods do not have the same

breadth in sequence space as synthetic pools, and precise control over the fixed regions not defined by primers is lost. Although not the best choice for creating a completely random library, mutagenesis is a valuable tool for exploring the mechanisms of nucleic acid catalysis. A cloned catalyst can be re-randomized and examined for changed activity. An evolving pool can be periodically mutagenized and reselected. This can, in theory, result in selection convergence on sequences around the most active catalytic motifs. The degree of mutation introduced at each position can be controlled by the type of mutagenic PCR employed. Chemical mutagenesis uses incorporation of nucleotide analogs to cause base-pair mismatching during PCR amplification and causes both transition and transversion mutations at an overall frequency of up to 19% per position per PCR [26]. Another technique uses natural nucleotides, Mn^{2+}, and an augmented Mg^{2+} concentration to generate a milder degree of mutagenesis of 10% per position per PCR [27, 28].

8.3.1.3 Standard PCR

PCR amplification of either the initial library or the products recovered from a cycle of *in vitro* selection should adhere to the same considerations as any PCR whose purpose is to produce high yields with low nonspecific amplification. Since the templates for a selection are usually relatively short, significant amplification can be achieved in a low number of cycles. Also the time for extension can be shortened to around 30 s. Thermostable polymerases have a fixed half-life under PCR conditions that is often a factor limiting the degree of amplification. Being able to shorten the total time of the PCR means that more flexibility is possible with respect to temperature without further reducing the activity of the polymerase. The annealing temperature can be raised to increase the specificity of priming, thereby reducing nonspecific products. The denaturation temperature can also be raised to ensure disruption of possible secondary structure. The ability to raise the temperature of these two steps is especially important when using long primers (~40 mer), and it also may be necessary to extend the length of the annealing step. PCR reactions should always be optimized for a particular primer set by performing pilot reactions from which an aliquot is taken at regular cycle intervals and analyzed by native PAGE and ethidium bromide staining (Section 8.3.1.5). The cycle number that gives the highest yield of product with the lowest amount of unwanted products should be used routinely.

There are various PCR protocols with associated degrees of fidelity. A low-fidelity, high-yield protocol using Taq DNA polymerase (NEB) for amplification of a template such as that in Figure 8.2 is as follows:

- 100 µL maximum reaction volume
- Buffer components at 1×:

KCl	10 mM
$(NH_4)_2SO_4$	10 mM
Tris-Cl (pH 8.8)	20 mM
Triton X-100	0.1% (v/v)

- Additional variable reaction components:

dsDNA template	~0.5 nM (quantitation not always necessary)
5′ primer	0.75 µM
3′ primer	0.75 µM
dNTPs	0.5 mM (each)
Mg^{2+}	2 mM
Taq Pol	0.025 U μL^{-1}

- Maximum of 40 cycles:

95 °C	for 1 min
55 °C	for 1 min
72 °C	for 30 s

The total yield should be ~50 pmol per 100 µL reaction after silica purification (Section 8.3.1.5).

An alternative procedure for high-fidelity amplification is achieved by lowering the dNTP and Mg^{2+} concentrations. In addition, a higher-fidelity polymerase such as Pfu^{-} (PfuUltra, Stratagene) can be used. Yields are generally reduced, but this procedure has specific applications in cloning or characterization of catalysts with known sequence, where mutations would impede analysis.

- 100 µL maximum reaction volume
- Buffer components at 1×:

KCl	10 mM
$(NH_4)_2SO_4$	10 mM
Tris-Cl [pH 8.8]	20 mM
Triton X-100	0.1% (v/v)

- Additional variable reaction components:

dsDNA template	~0.5 nM (quantitation not always necessary)
5′ primer	0.75 µM
3′ primer	0.75 µM
dNTPs	0.05 mM (each)
Mg^{2+}	1 mM
Taq Pol	0.025 U μL^{-1}

- Hot start:

95 °C	for 2 min

- Maximum of 25 cycles:

95 °C	for 30 s
55 °C	for 1 min
72 °C	for 30 s

Pfu DNA polymerase can be substituted for Taq DNA polymerase, but the manufacturer (Stratagene) recommends using the supplied buffer.

Total yield should be ~35 pmol per 100 µL reaction after purification (Section 8.3.1.5).

8.3.1.4 Mutagenic PCR

As discussed in Section 8.3.1.2, mutagenic PCR has various applications for producing starting libraries and remutagenizing evolved pools of catalysts. However, chemical synthesis is much more practical than even high-error-rate PCR procedures for preparing randomized sequence pools. These chemical mutagenesis methods have found most use for randomizing very large templates. A mild PCR mutagenesis procedure using near-normal conditions has proven to be useful for *in vitro* selection when remutagenizing an evolved pool is of interest.

- 100 µL maximum reaction volume
- Buffer components at 1×:

KCl	50 mM
gelatin	0.01% (w/v)
Tris-Cl (pH 8.3)	10 mM

- Additional components:

dsDNA template	0.2 nM
5′ primer	0.5 µM
3′ primer	0.5 µM
dATP	0.2 mM
dCTP	1 mM
dGTP	0.2 mM
TTP	1 mM
$MgCl_2$	7 mM
Taq Pol	0.05 U µL-1
$MnCl_2$	0.5 mM (made fresh and added last)

- Maximum of 40 cycles:

95 °C	for 30 s
53 °C	for 1 min
72 °C	for 1 min

Total yield should be ∼10 pmol of proper-length (∼150 bp) product per 100 µL reaction.

Native polyacrylamide gel purification (Section 8.3.1.5) is required to isolate products of the appropriate size.

If further mutagenesis is required, fractions of the gel-purified product can be used as templates for multiple mutagenic PCR reactions that are later pooled. This can even be repeated a third time, if the product is gel-purified again after the last PCR. Higher yields than obtained from gel purification are usually needed. If so, the final mutagenized pool can be amplified under higher-yield, nonmutagenic conditions and subsequently purified by high-salt adsorption to a silica membrane (Section 8.3.1.5).

8.3.1.5 Purification of Libraries

Gel purification is a rapid and efficient way of isolating nucleic acids of the appropriate size from syntheses, PCR reactions, ligations, or tethered product-binding reactions. For preparative separation of random libraries (~150 bases) the following two types of polyacrylamide gels are used:

- Native:
  1.5 mm × 16 cm × 14 cm
  6% acrylamide (19:1) in 1× TBE
  native loading buffer
  The gel should be run for 2 h at 8 W, and results in separation of dsDNA or folded RNA in a range of ~250–50 base pairs.
- Denaturing:
  1.5 mm × 16 cm × 14 cm
  6% acrylamide (19:1) in 1× TBE
  8.3 M urea
  denaturing loading buffer
  samples heated to 72 °C for 5 min
  gel preheated for 30 min, then kept above room temperature
  Run the gel for 3 h at 8 W to separate RNA or ssDNA in a range of ~500–80 bases.

Nucleic acids can be visualized by ethidium bromide staining, UV shadowing, or phosphorimaging of radioactive samples. A sterile scalpel should be used to excise the separated product which can then be eluted by electrophoresis (1× Tris-borate, pH 8.3) into a 30 kDa molecular weight cutoff (MWCO) filter (Millipore). The product is concentrated by centrifugation, dialyzed, and resuspended in the buffer of choice. The yield of nucleic acid is typically 75%, and ethanol precipitation is not needed.

Passive elution is another commonly used technique for isolation of RNA from polyacrylamide gels. The steps for this purification method, also called 'freeze-squeeze', are as follows:

1. Place the gel slice directly (do not allow it to dry) into a 1.7-mL microcentifuge tube.
2. Freeze the gel slice at −20 °C overnight or in a dry ice/acetone bath for 10 min.
3. Crush (homogenize) the frozen gel slice with an appropriate device. A syringe plunger of appropriate size works well.
4. Add 700 µL of extraction buffer (400 mM NaCl and 2 mM EDTA) to the homogenized gel slice.
5. Vortex vigorously for 1 min.
6. Centrifuge the slurry for 3 min at 10 000 × g and save the supernatant.
7. Repeat steps 4–6 for a total of 4 extractions.
8. Filter the combined extracts (2.8 mL) through a 0.45-µm spin filter (Corning).

9. Concentrate the filtrate on a MWCO filter (30 kDa).
10. Wash the concentrate three times on the filter with 500 µL H_2O and collect the retentate for quantitation and use.

Yields from this procedure are typically 30%–50%.

Another method of purification, especially of PCR and transcription products, is by adsorption of nucleic acids to silica under denaturing, high-salt conditions. This method should be used with caution when performed as part of a selection using modified RNA, since it is possible to evolve catalytic RNA sequences that stick tightly to silica or, even worse, not at all. Nevertheless, for evolved RNA pools, washing with denaturing, high-salt buffer removes ethidium bromide, protein, and nucleotides rapidly, and subsequent elution with low-salt buffer or H_2O typically provides the desired oligonucleotides in high yield and purity. Nucleic acids entering into an *in vitro* selection should be purified in this or some other manner to remove such contaminants. Several companies (e. g., Qiagen) sell products based on this process, which are extensively used for high-throughput screening (HTS) applications.

8.3.2 Transcription

An important feature of RNA *in vitro* selections is T7 RNA polymerase. Many different functionalized modified ribonucleotides can be efficiently incorporated into the transcripts with high fidelity by T7 RNA polymerase. The same may be true for incorporation of modified deoxyribonucleotides into DNA using DNA polymerases, but the ability to incorporate a wide variety of modified nucleotides has not been fully investigated. Although some RNA catalysts have been discovered that contain only natural nucleotides, the scope of catalysis can be greatly increased by adding functional groups to the nucleotide bases, so that they add conformational flexibility, nucleophilicity, electrophilicity, metal coordination sites, and acid/base characteristics. Some of the various modified nucleotides that can be incorporated to an extent comparable to that of the analogous natural nucleotides are illustrated in Figure 8.3 (Vaught, Dewey, Eaton unpublished results). These modifications are well tolerated by T7 RNA polymerase with regard to sequence diversity and overall fidelity. When using any new modified nucleotide for the first time, it is important to check the kinetics and fidelity of incorporation during transcription compared to that of the natural nucleotide by using a gel-fidelity assay [29]. It is also important to determine the base composition by nuclease digestion followed by HPLC analysis [6].

Chemical post-transcriptional modifications are possible depending on the design of the original template. For example, by altering the transcription conditions to favor incorporation of a particular terminal 5′ nucleotide or modified analog, chemical conjugation can be used to tether a potential substrate. Many variations on this theme can be employed. Modified nucleotides can be coupled to a potential substrate by organic synthesis and then incorporated into an oligonucleotide

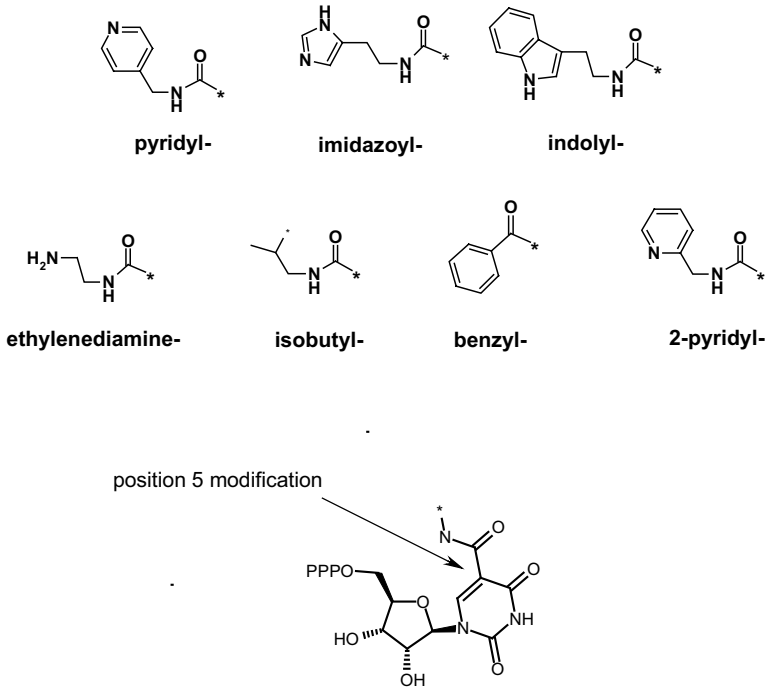


Fig. 8.3. During transcription by T7 RNA polymerase, uridine triphosphates with the modifications shown can be incorporated into the RNA transcript without loss of sequence information or diversity.

by either phosphoramidite oligonucleotide synthesis or enzymatic incorporation during transcription [7, 21, 30]. An elegant example of 5′-end modification is the incorporation of GMP (or GMPS) as the 5′ ribonucleotide. This establishes a single terminal phosphate group allowing for ligation by T4 DNA ligase to chemically modified oligomers. This ligation procedure (Section 8.3.3) allows for the creation of diverse RNA sequence pools consisting of both internally (from transcription) and end-modified products. A typical procedure for the preparation of 5′-(4-pyridylmethyl)UTP [31, 32] modified transcript (Figure 8.3) from a random DNA library (Section 8.3.1.1) that establishes a point of further modification by ligation (Section 8.3.3) is as follows:

- 100 µL total reaction volume
- 30 pmol purified DNA template library (Section 8.3.1.5)
- ATP, CTP, GTP 1 mM each
- 5-(4-pyridylmethyl)UTP 1 mM (Figure 8.3)
- GMP 30 mM
- [α-^{32}P] ATP 10–30 µCi per 100 µL reaction volume
- Nuclease-free H_2O (up to final volume of 100 µL)

- Transcription buffer components at 1×:

Tris (pH 7.9)	40 mM
$MgCl_2$	6 mM
spermidine	2 mM
NaCl	10 mM
DTT (optional)	10 mM

1. Add above components, vortex, centrifuge briefly.
2. Add the following enzymes:

RNasin (Promega)	$1\,U\,\mu L^{-1}$
T7 RNA Pol (Promega)	$3.84\,U\,\mu L^{-1}$

3. Mix gently, centrifuge briefly.
4. Incubate reaction mixture for 5 h at 37 °C.

   Addition of inorganic pyrophosphatase at $0.025\,U\,\mu L^{-1}$ may increase yields by hydrolyzing pyrophosphate, thereby decreasing inhibition of T7 by pyrophosphate, but this is an optional component.

   If DNA carryover is a concern, you may add RNase-free DNase before the next step:

5. Purify modified RNA transcript as described in Section 8.3.1.5.
6. Quantitate 2 µL of eluate by Cherenkov counting (using the ratio of [α-^{32}P] ATP/ATP)

   Yields should be 0.5–1 nmol of body-labeled transcripts.

   All previous and subsequent steps of the RNA selection process should be performed following strict RNase-free conditions. This includes the use of DEPC-treated H_2O for all solutions and the use of aerosol barrier pipette tips. Any glassware used (including electrophoresis and electroelution apparatus) should be treated with an RNase inhibitor spray (e. g., RNase AWAY, Molecular BioProducts).

8.3.3 Ligation

As discussed in Section 8.3.2, there are several ways for substrates to be directly attached to the 5′ end of RNA by incorporation of a modified ribonucleotide monophosphate during transcription. A more flexible system can be employed that allows for the use of untemplated or nonintercalating substrates. Using a highly water-soluble, conformationally flexible linker between the 5′-end of the RNA and the substrate can allow the substrate sufficient freedom to move to the catalytic active site, even if that site is far removed from the 5′ end of the RNA. The most commonly used linker is poly(ethylene glycol) (PEG). This relatively inert linear polymer is soluble in the aqueous buffers used for many RNA catalyzed reactions and can improve the solubility of substrates. Typically, PEG substrates are converted to phosphoramidite regents that can be used in solid-phase synthesis of

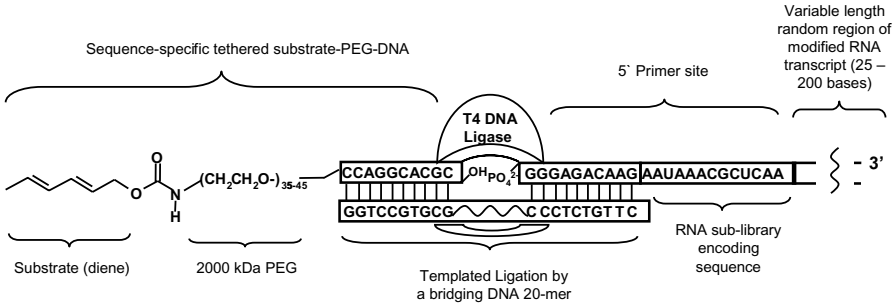


Fig. 8.4. Using a bridged ligation to add a PEG-tethered reactant to the already-modified RNA transcript creates a diverse pool of potential catalysts with tethered reactants.

5′-end-modified DNA [7, 21, 22]. The modified DNA can then be attached to the 5′ end of RNA libraries by ligation.

For example, an RNA/DNA ligation procedure is shown in Figure 8.4, in which an RNA transcript is attached to a 10-mer DNA oligonucleotide tethered to a substrate through a PEG linker. A substrate tethered by a 2000-MW PEG linker to a catalyst could have access to the entire RNA surface in such an intramolecular system (Figure 8.4). In this example, the substrate has an estimated concentration of approximately 150 μM relative to its attached RNA. A further advancement possible with this substrate–PEG–DNA–RNA construct is that the specific ligation sequence used at the 5′ end of the RNA is inherently encoded with information about which substrate is attached or (better yet) has reacted in a cycle of selection. This feature allows for multiple substrates to be tested simultaneously in the same RNA population. The attached substrate 'winners' of the selection can be tracked by quantitative PCR, providing valuable information about the chemistry that is occurring in concert with the evolving catalysts. The example given in Figure 8.4 is a generalized example of a diene–PEG–DNA–RNA ligation product that catalyzes a [4 + 2] cycloaddition (Section 8.3.4) with an activated dienophile. The procedure for its production is given in Section 8.3.3.1.

8.3.3.1 Ligation Procedure

All reagents for this step must be of the highest purity. Purification of the transcript library is described above (Section 8.3.1.5). The 10-mer-PEG substrate must be synthesized, purified by HPLC (PRP-1 column, Hamilton), and characterized by mass spectrometry before use. Additionally, each of these components should be accurately quantified by absorbance at 260 nm, using background subtraction at 320 nm. The 20-mer DNA bridge can be ordered from a commercial DNA synthesis facility (Operon). Ratios of the components in this ligation procedure are critical. Nucleic acids entering into this step should be in H_2O, because excess salt can inhibit T4 DNA ligase [33]. T4 DNA ligase requires ATP, which is included in

the buffers used. As a general rule, all buffers, ligation components, enzymes, and nucleotides should be stored at −20 °C.

- 100 µL reaction volume
- T4 reaction buffer components at 1×:

Tris-Cl (pH 7.8)	30 mM
Mg^{2+}	10 mM
DTT	10 mM
ATP	10 mM

- T4 storage/dilution buffer:

Tris-Cl (pH 7.5)	20 mM
KCl	60 mM
DTT	5 mM
EDTA	1 mM
glycerol	50% (v/v)

- Nuclease-free ddH_2O
- T4 DNA ligase (Promega) 0.317 Weiss U μL^{-1}
- RNasin (Promega) 1 U μL^{-1}
- Modified, purified transcript of RNA
- HPLC-purified 10-mer-PEG substrate
- 20-mer bridging oligonucleotide

1. Mix RNA library (5 µM), substrate-PEG-DNA (10 µM), and 20-mer DNA bridge (15 µM) from such stock concentrations that the volume of the mixture is ≈50 µL.
2. Mix thoroughly by vortexing, centrifuge briefly.
3. Incubate mixture at 70 °C for 5 min.
4. Allow the sample to cool to room temperature while performing steps 5–6.
5. Mix the reaction buffer (1×) with the appropriate volume of water to have a final reaction volume of 100 µL.
6. (Optional) Depending on the concentration of T4 DNA ligase stock, you may need to supplement up to 0.1% (v/v) of the reaction mixture with the storage/dilution buffer.
7. Vortex the buffer(s) and water mix and centrifuge briefly,
8. Add RNasin (1 U μL^{-1}) to the buffer/water solution and mix gently.
9. Add T4 DNA ligase (0.317 U μL^{-1}) to the buffer/water/RNasin mix, mix gently, and centrifuge briefly.
10. Add this buffer/water/RNasin/T4 mix to the annealed ligation components and subject it to the following thermocycler sequence:

37 °C	for 3 h
22 °C	for 3 h
17 °C	for 3 h
4 °C	hold (overnight)

A master mix can be made from the final mixture produced from steps 5–9 and dispensed into many ligation reactions.

8.3.3.2 Purification of the Ligation Product

As described in Section 8.3.1.5, purification of ligation products is most easily achieved by gel-shift PAGE. The additional molecular weight of the ligated modification is enough (e. g., 2000 MW PEG plus 10-mer DNA) to separate it from the unligated RNA by preparative 6% denaturing PAGE (Section 8.3.1.5). Usually the sample is radioactive and can therefore be excised by laying the gel on an actual-size phosphorimage paper template. The gel slice should not be allowed to dry before elution. The gel slice can be immediately placed in a 30 kDa MWCO filter that fits directly into an electroelution apparatus (Millipore), which can efficiently recover ~75% of product by applying ~200 V for 3 h. The product can then be directly concentrated by centrifugation, washed extensively on the same filter, and quantified by Cherenkov counting. It is important that the electroelution buffer (Section 8.2) does not contain any EDTA, because this chelator, if carried forward into the catalytic reaction step, may sequester important metals and inhibit the reaction.

8.3.4 Nucleic Acid-catalyzed Reactions

Nucleic acids catalyze many different types of reactions. Some RNA-catalyzed transformations show stereoselectivity [10, 34]. The potential scope of organic reactions is quite broad, with a commensurate variability in reaction conditions. The essential components present in successful nucleic acid-catalyzed reactions are divalent metal ions such as Mg^{2+}, Ca^{2+}, Cu^{2+}, Zn^{2+}, as well as K^{+} [7, 10, 21, 35, 36]. A buffer is also required but should not contain functional groups that are reactive under the reaction conditions. A commonly used buffer is HEPES (2-[4-(2-hydroxyethyl)-1-piperazine]ethanesulfonic acid). These essential components are present to maintain the RNA's tertiary structure and prevent its aggregation. Because these reactions are carried out in aqueous solution, the addition of a co-solvent (e. g., DMSO or EtOH) may be necessary, depending on the solubility of the substrates.

Additional reaction components can be used that may enhance or be essential for catalytic activity in a particular system. These components may serve functions analogous to the cofactors used by protein catalysts (e. g., ATP, NADH, or metal ions). These components can be supplied free in solution or incorporated into the RNA library as previously described. As an example, Figure 8.5 outlines the RNA-catalyzed carbon-carbon bond-forming [4 + 2] cycloaddition reaction between a tethered diene substrate, (2E.4E)-hexa-2,4-dien-1-O-PEG (**1**), and 1-biotinamido-4-[4′-(maleimidomethyl)cyclohexanecarboxamido] butane (BMCC-biotin, **2**), a dienophile that is free in solution. The RNA catalyzes the formation of (**3**), which contains biotin for partitioning purposes.

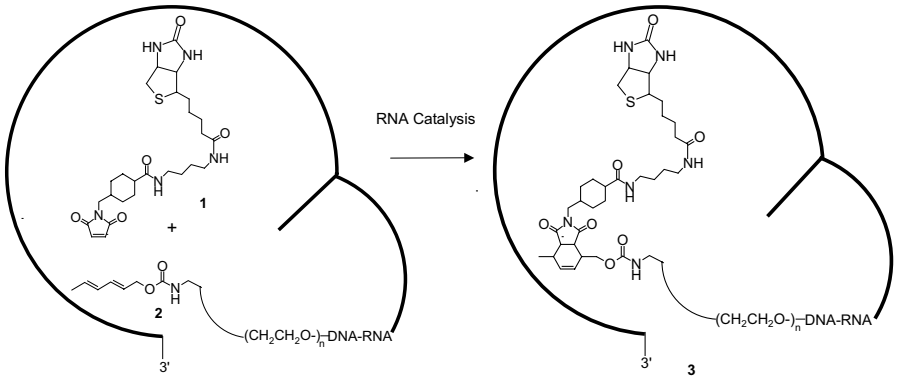


Fig. 8.5. Specific example of an RNA catalyzed reaction: [4 + 2] Diels-Alder cycloaddition between an RNA-tethered diene and a biotin-linked maleimide dienophile (BMCC-biotin). Cycloaddition creates an affinity-tagged (biotin) RNA product that can be captured and purified.

8.3.4.1 Model Reaction Conditions

- 100 µL total volume
- Reaction buffer at 1×:

HEPES (pH 7.0)	50 mM
NaCl	200 mM
KCl	200 mM

- Additional (variable) components:

Ca^{2+}, Mg^{2+}	1 mM
Al^{3+}, Ga^{2+}, Mn^{2+}, Fe^{2+}, Co^{2+}, Ni^{2+}, Zn^{2+}, Cu^{2+}	10 µM

(Metal-ion solutions are made as fresh as possible and stored under argon at −20 °C.)

ethanol	10% (v/v)
DMSO	2% (v/v)

- BMCC-biotin made prior to use as stock solution of 10 mM in DMSO
- Ligated, modified RNA construct supplied in H_2O (from electroelution)

1. Mix the reaction buffer and metal ions by vortexing.
2. Add the RNA construct to a final concentration of 500 nM.
3. Add ethanol and vortex the entire mixture.
4. Add BMCC-biotin to a final concentration of 100 µM (its addition accounts for the total amount of 2% DMSO in the reaction mix)
5. Bring the final volume to 100 µL with H_2O and vortex.
6. Allow the reaction to proceed at 25 °C for 1.5 h.
7. Dilute the reaction mixture by adding 4 volumes of H_2O. Remove unreacted BMCC-biotin through a 30 kDa MWCO filter (Millipore) by centrifugation for 8 min at 11 000 × g. Repeat this step to ensure that all unreacted BMCC has been removed before proceeding to the partitioning step (Section 8.3.6).

8.3.5 Reverse Transcription

After the RNA pools have been enriched for catalysts, it is essential to amplify these RNA sequences so that additional cycles of selection can be performed. To accomplish this, the RNA pool at each selection cycle must be reverse-transcribed to give cDNA that can subsequently be taken into PCR where significant amplification can be achieved. A typical reverse transcription is as follows:

- Total volume of 100 µL
- Buffer components at 1×:

Tris-Cl (pH 8.3)	50 mM
KCl	75 mM
$MgCl_2$	5 mM

- Reaction components:

RNA construct	not quantified (50 fmol) in H_2O
dNTPs	0.75 mM each
3′ primer	1 µM
RNasin (Promega)	1 U µL^{-1}
Superscript II (Invitrogen)	10 (U µL^{-1})

1. Mix buffer components, dNTPs, primer, and H_2O (to bring to final volume) by vortexing.
2. Add RNA construct (variable volume) and vortex.
3. Add RNasin and Superscript II, mix gently, and centrifuge.
4. Incubate at 42 °C for 45 min.

8.3.5.1 Nucleic Acid Purification after Reverse Transcription

Proteins have high affinity to specific forms of silica (Strataclean, Stratagene), whereas under the appropriate conditions nucleic acids do not. This activated silica can be washed and added directly to the terminated reverse-transcription reaction mix to efficiently adsorb proteins (enzymes), leaving a nucleic acid-enriched supernatant:

Cleanup of a 100-µL Reverse Transcription Reaction

- Wash buffer:

Tris-Cl (pH 8.5)	25 mM
KCl	5 mM
$MgCl_2$	5 mM
NaCl	75 mM
Triton X-100	0.05% (v/v)

1. Equilibrate the silica by washing with a low-salt buffer containing a nonionic detergent (wash buffer). The silica is supplied as a 50% slurry and must be thoroughly resuspended before use.

2. Transfer 25 µL of the homogenous slurry to a 0.5-mL microcentrifuge tube and centrifuge for 2 min at 8000 × *g*.
3. Discard the supernatant and replace it with 400 µL of wash buffer.
4. Vortex this slurry for ∼30 s, centrifuge again for 2 min at 8000 × *g* and discard the supernatant.
5. Repeat steps 3–4 for a total of 4 washes.
6. Resuspend the final washed pellet back to a 50% slurry in wash buffer (add 12.5 µL).
7. Add the reverse transcription reaction mix (100 µL) directly to the tube containing the washed 25 µL slurry of silica, vortex briefly, and incubate at 25 °C with constant inversion for 10 min.
8. Centrifuge this mixture as above and carefully transfer the supernatant to a new tube.
9. Filter the final supernatant through a 0.22 µm spin filter (Spin-X, Corning).

Recovery should be ∼90% from input into the reaction through this purification procedure (∼45 pmol total yield), as quantified by Cherenkov counting.

At this point, the RNA constructs have been purified away from any protein and small reactant molecules. To further dialyze the sample, it can be diluted and concentrated once again on a 30 kDa MWCO filter and recovered in the desired buffer. This should leave pure modified nucleic acids ready to enter the affinity-capture stage of the *in vitro* selection – partitioning.

8.3.6 Partitioning

Selective capturing of the active nucleic acid sequences is a crucial factor in determining the success of an RNA catalyst *in vitro* selection. A convenient technique for partitioning active from inactive RNA sequences involves capture of biotin by streptavidin. Biotin is a widely used affinity tag because of its strong binding ($K_d \sim 10^{-15}$) to streptavidin. Potential substrates are typically synthesized with biotin tags so that the products formed contain biotin, and therefore the active RNA catalysts can be isolated by binding to streptavidin. Streptavidin can be attached to an insoluble resin or other solid surfaces to allow for easy partitioning (Figure 8.6). In addition, streptavidin-biotin binding can be done in solution and the products can be analyzed and partitioned by 6% denaturing PAGE (Section 8.3.1.5). This solution-phase partitioning method has the advantage that PAGE analysis is performed as part of the protocol, as opposed to simply counting what radioactivity is retained on a solid phase. The streptavidin-dependent gel-shifted product can be quantified by phosphorimaging or scintillation counting of excised gel bands. Phosphorimaging of PAGE gel-shifted bands has also been used to obtain kinetic information [7]. Alternative methods involving HPLC partitioning of streptavidin-captured biotinylated products and quantitation by UV have also proved useful for substrates with good chromophores [22].

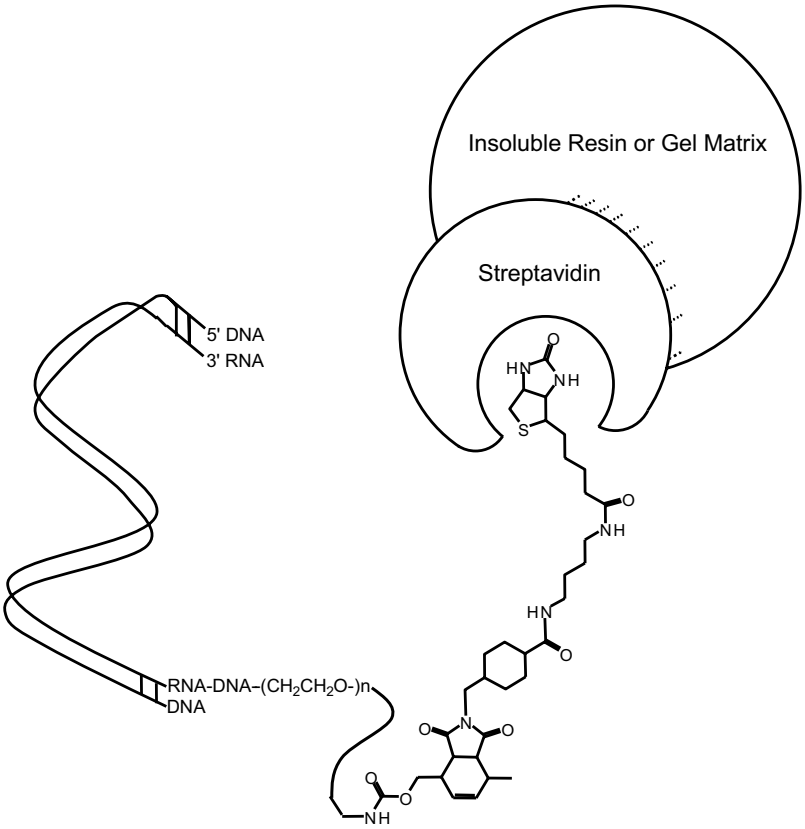


Fig. 8.6. By utilizing a streptavidin-coupled immobilized matrix or gel-shift PAGE, the affinity tag biotin allows the RNA catalysts to be captured and separated from the noncatalytic RNA.

8.3.6.1 Gel-shift Protocol for Partitioning and Analysis

1. Make a fresh solution of 0.1 mM streptavidin in 50 mM HEPES (pH 7.2).
2. Prepare a 25 µM solution of recovered RNA constructs (Figure 8.5) in 10 µL of H_2O.
3. Add 10 µL of streptavidin stock to the 25 µM RNA solution.
   - Final concentrations:

RNA constructs	12.5 µM
streptavidin	25 µM
4. Allow the mixture to incubate for 30 min.
5. During this 30 min period, prepare and warm a gel for 6% denaturing PAGE analysis.
6. Mix incubated samples with 10 µL of 3× concentrated formamide loading buffer.

7. Heat samples to 72 °C for 5 min., immediately load onto the previously prepared gel, and apply running conditions as in Section 8.3.1.5.

   Phosphorimaging quantitation is recommended.

   Streptavidin-bound inhibition of product migration is clearly visible, and the bands can be excised and purified as described in Section 8.3.1.5.

8. Recovered products are now ready to start the cycle again with PCR amplification of the cDNA/RNA templates.

Proper controls should be included (e. g., no streptavidin, no substrate, and an RNA construct with an absent or inert tethered substrate) to prove that the reaction involves the linked substrate. Also, it is highly advisable to save some PCR product from each cycle of selection. These saved samples can be used to check for convergence of sequences into families and can also help avoid having to start the selection from the beginning.

8.3.6.2 Other Methods of Partitioning

RNA catalysis and *in vitro* selection are ever increasing in scope, and the method presented in Section 8.3.6.1 is by no means the only alternative for separating reacted/active-catalyst complexes. Most research groups have used this type of partitioning procedure, based on some type of biotin-product capture by streptavidin. Other partitioning methods are possible and this step in the overall RNA catalysis selection cycle is where many new innovations need to occur to advance the field.

The method of RNA catalyst capture really depends on the properties of the tethered product. Another method for selective partitioning is by chromatographic (HPLC) methods based on the properties of products that do not contain an affinity tag such as biotin. If the overall electrostatic or steric properties of the RNA product are significantly changed so that they are much different than those of the unreacted RNA, chromatographic methods can be used to separate and recover the products and obtain their sequence information [14].

Cleavage of the captured RNA product is an effective partitioning alternative. During partitioning, after affinity-capture and washing, nucleic acids can be cleaved from the bound product [23, 35]. The bound product and related sequences can be taken forward separately for analysis. This greatly reduces contamination of the recovered pool by sequences that did not catalyze a reaction with the tethered substrate. The design of a successful selection benefits from carefully considering the multitude of partitioning parameters that can be established and changed during the course of a selection.

8.3.6.3 Sequencing and Cloning

Sequencing of pools recovered as PCR products from regular cycles of selection can be performed and compared to the starting RNA pool. This provides information

about the convergence of sequence information in the evolving pool. This 'bulk' sequencing provides an indication of how the selection is proceeding. It is most easily performed by dideoxy cycle-sequencing with thermosequenase (Amersham) and a ^{32}P end-labeled primer.

At the end of a selection experiment, the individual 'winning' sequences can be spatially isolated from the evolved pool by standard cloning procedures. The amplified DNA pool from any cycle of selection can be treated as a 'pure' solution of insert and cloned in a variety of ways, most typically by blunt-ended ligation, into any variety of plasmid vectors (e. g., PCR-Script, Stratagene). Bacteria can be transformed, plated, isolated, and plasmid-purified by standard protocols or high-throughput facilities (GRL, North Carolina State University). High-throughput sequencing (recommended) can then be performed on the individual catalysts. Sequence information can be used to align RNA catalysts and group them into families with conserved motifs. The individual sequences can be transcribed, ligated, and used under various conditions, as described, to determine their catalytic activity [11, 37].

8.4 Troubleshooting

Most RNA *in vitro* selection experiments have been designed to find aptamers. Indeed, if reactions are not carefully monitored, RNA *in vitro* selection for catalysts can be overwhelmed by aptamers. Additional steps, such as a negative selection, can be included in the selection procedure to help mitigate unwanted aptamer selection. In the selection outlined in Section 8.3, an additional step can be added, in which the RNA pool is exposed to streptavidin before the reaction step. The streptavidin-RNA aptamer complexes can be removed by various methods (e. g., MWCO filtration, gel purification, activated silica partitioning), and the remainder of the RNA population carried forward into the reaction step and partitioned for RNA catalysis. This example is one of negative selection for streptavidin aptamers. The same sort of negative selection can be applied to the inert matrix used for product capture in the partitioning step or for aberrant PAGE gel-shift bands that are not the desired catalysts.

Another issue that should be considered carefully is the purity of nucleic acids used at each step. Since it is known that RNA can self-cleave and ligate, DNA produced from PCR that is entering the next cycle of selection should be size-purified by native PAGE (Section 8.3.1.5). Finally, *all* solutions should be prepared from DNase/RNase-free reagents in DEPC H_2O, 0.2 µm filtered, and stored at 4 °C (buffers) or −20 °C (especially nucleotide solutions).

8.5 Major Applications

It is tempting to speculate on the potential role of RNA as a catalyst for carrying out organic reactions. In particular, it might be of interest to use RNA to prepare highly functionalized chiral molecules with biological activity [38]. Creating highly efficient RNA catalysts for organic synthesis is a goal that has recently received much attention. Perhaps, as we learn more about how to improve these new biocatalysts, they can become competitive with more traditional catalysts, especially for high value-added pharmaceuticals and synthetic intermediates.

RNA catalysis has been proposed for use in preparing combinatorial libraries of organic structures for drug discovery [39]. As we learn more about the scope, reactivity, and specificity of RNA as a catalyst for organic reactions, it should be possible to use RNA to create new chemical diversity that parallels that found in biological systems, where proteins are the catalysts in the formation of natural products.

References

1. Tuerk, C., Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage-T4 DNA polymerase. *Science 249*, 505–510 (**1990**).
2. Ellington, A.D., Szostak, J.W. In vitro selection of RNA molecules that bind specific ligands. *Nature 346*, 818–822 (**1990**).
3. Tarasow, T.M., Eaton, B.E. Dressed for success: realizing the catalytic potential of RNA. *Biopolymers 48*, 29–37 (**1998**).
4. Lee, S.E., Sidorov, A., Gourlain, T., Mignet, N., Thorpe, S.J., Brazier, J.A., Dickman, M.J., Hornby, D.P., Grasby, J.A., Williams, D.M. Enhancing the catalytic repertoire of nucleic acids: simultaneous incorporation of amino and imidazolyl functionalities by two modified triphosphates during PCR. *Nucleic Acids Res. 29*, 1898–1905 (**2001**).
5. Vaish, N.K., Fraley, A.W., Szostak, J.W., McLaughlin, L.W. Expanding the structural and functional diversity of RNA: analog uridine triphosphates as candidates for in vitro selection of nucleic acids. *Nucleic Acids Res. 28*, 3316–3322 (**2000**).
6. Dewey, T.M., Mundt, A.A., Crouch, G.J., Zyzniewski, M.C., Eaton, B.E. New uridine derivatives for systematic evolution of RNA ligands by exponential enrichment. *J. Am. Chem. Soc. 117*, 8474–8475 (**1995**).
7. Tarasow, T.M., Tarasow, S.L., Eaton, B.E. RNA-catalysed carbon-carbon bond formation. *Nature 389*, 54–57 (**1997**).
8. Ordoukhanian, P., Joyce, G.F. RNA-cleaving DNA enzymes with altered regio- or enantioselectivity. *J. Am. Chem. Soc. 124*, 12499–12506 (**2002**).
9. Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E., Chech, T.R. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell 31*, 147–157 (**1982**).

10. Nieuwlandt, D., West, M., Cheng, X., Kirshenheuter, G, Eaton, B.E. The first example of an RNA urea synthase: selection through the enzyme active site of human neutophile elastase. *ChemBioChem 4*, 649–662 (**2003**).
11. Stuhlmann, F., Jaschke, A. Characterization of an RNA active site: interactions between a Diels-Alderase ribozyme and its substrates and products. *J. Am. Chem. Soc. 124*, 3238–3244 (**2002**).
12. Zhang, B.L., Cech, T.R. Peptide bond formation by in vitro selected ribozymes. *Nature 390*, 96–100 (**1997**).
13. Jadhav, V.R., Yarus, M. Acyl-CoAs from coenzyme ribozymes. *Biochemistry 41*, 723–729 (**2002**).
14. Illangasekare, M., Sanchez, G., Nickles, T., Yarus, M. Aminoacyl-RNA synthesis catalyzed by an RNA. *Science 267*, 643–647 (**1995**).
15. Breaker, R.R., Joyce, G.F. A DNA enzyme that cleaves RNA. *Chem. Biol. 1*, 223–229 (**1994**).
16. Li, Y.F., Sen, D. A catalytic DNA for porphyrin metallation. *Nat. Struct. Biol. 3*, 743–747 (**1996**).
17. Joyce, G.F. Nucleic acid enzymes: playing with a fuller deck. *Proc. Natl. Acad. Sci. USA 95*, 5845–5847 (**1998**).
18. Fitzwater, T., Polisky, B. A SELEX primer. *Meth. Enzymol. 249*, 275–301 (**1990**).
19. Zaccolo, M., Williams, D.M., Brown, D.M., Gherardi, E. An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J. Mol. Bio. 255*, 589–603 (**1996**).
20. Suga, H., Lohse, P.A., Szostak, J.W. Structural and kinetic characterization of an acyl transferase ribozyme. *J. Am. Chem. Soc. 120*, 1151–1156 (**1998**).
21. Wiegand, T.W., Janssen, R.C., Eaton, B.E. Selection of RNA amide synthases. *Chem. Biol. 4*, 675–683 (**1997**).
22. Seelig, B., Jaschke, A. A small catalytic RNA motif with Diels–Alderase activity. *Chem. Biol. 6*, 167–176 (**1999**).
23. Hausch, F., Jaschke, A. Libraries of multifunctional RNA conjugates for the selection of new RNA catalysts. *Bioconjugate Chem. 8*, 885–890 (**1997**).
24. Ciesiolka, J.M.I.I.M., Nickles, T., Welch, M., Yarus, M., Zinnen, S. Affinity selection–amplification from randomized ribooligonucleotide pools. *Meth. Enzymol. 267*, 315–335 (**1996**).
25. Nieuwlandt, D., Kellogg, E., Wecker, M., Qui, J., Wolk, S., Tarasow, T., Dewey, T., Eaton, B. Anti-MRSA drug leads from evolutionary chemistry. American Society of Microbiology (ASM) Annual Interscience Conference on Antimicrobial Agents and Chemotherapy (ICAAC). 12–17 (**2001**).
26. Zaccolo, M., Williams, D.M., Brown, D.M., Gherardi, E. An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J. Mol. Bio. 255*, 589–603 (**1996**).
27. Vartanian, J.P., Henry, M., Hobson, W.S. Hypermutagenic PCR involving all four transitions and a sizeable proportion of transversions. *Nucleic Acids Res. 24*, 2627–2631 (**1996**).
28. Cadwell, R.C., Joyce, G.F. Mutagenic PCR. *PCR Methods Appl. 3*, S136–S140 (**1994**).
29. Boyer, J.C., Bebenek, K., Kunkel, T.A. Analyzing the fidelity of reverse transcription and transcription. *Meth. Enzymol. 275*, 523–537 (**1996**).
30. Seelig, B., Jaschke, A. Ternary conjugates of guanosine monophosphate as initiator nucleotides for the enzymatic synthesis of 5′-modified RNAs. *Bioconjugate Chem. 10*, 371–378 (**1999**).

31. Dewey, T. M., Zyzniewski, C., Eaton, B. E. The RNA world: functional diversity in a nucleoside by carboxamidation of uridine. *Nucleosides Nucleotides 15*, 1611–1617 (**1996**).
32. Eckstein, L.J. Rapid and efficient synthesis of nucleoside 5′-O-(1-thiotriphosphates), 5′-triphosphates and 2′,3′-cyclophosphorotioates using 2-chloro-4H-1,3,2-benzodioxaphosphorin-4one. *J. Org. Chem. 54*, 631–635 (**1989**).
33. Rossi, R., Montecucco, A., Ciarrocchi, G., Biamonti, G. Functional characterization of the T4 DNA ligase: a new insight into the mechanism of action. *Nucleic Acids Res. 25*, 2106–2113 (**1997**).
34. Seelig, B., Keiper, S., Stuhlmann, F., Jaschke, A. Enantioselective ribozyme catalysis of a bimolecular cycloaddition reaction. *Angew. Chem.. Int. Ed. Engl. 39*, 4576–4579 (**2000**).
35. Sengle, G., Eisenfuhr, A., Arora, P.S., Nowick, J.S., Famulok, M. Novel RNA catalysts for the Michael reaction. *Chem. Biol. 8*, 459–473 (**2001**).
36. Illangasekare, M., Yarus, M. Specific, rapid synthesis of Phe-RNA by RNA. *Proc. Natl. Acad. Sci. USA 96*, 5470–5475 (**1999**).
37. Tarasow, T.M., Tarasow, S.L., Tu, C., Kellogg, E., Eaton, B.E. Characteristics of an RNA Diels-Alderase active site. *J. Am. Chem. Soc. 121*, 3614–3617 (**1999**).
38. Jaschke, A. In vitro selected oligonucleotides as tools in organic chemistry. *Synlett* **6**, 825–833 (**1999**).
39. Dewey, T.M., Nieuwlandt, D., Tarasow, T. Integrated drug discovery in a test tube. *Current Drug Discovery* **July** 21–25 (**2002**)..

9 High-throughput Screening of Enantioselective Industrial Biocatalysts

Manfred T. Reetz

9.1 Introduction

The catalytic asymmetric synthesis of enantiomerically pure or enriched organic compounds is of considerable academic and industrial interest [1]. For example, the so-called 'chiral market' in the area of pharmaceutical products currently exceeds $100 billion per year [2]. Many of the intermediate chiral compounds needed for the synthesis of the final therapeutic drug are prepared in the laboratories of organic chemists. The same applies to the production of chiral plant-protective agents, fragrances, and other products. The two major options for asymmetric catalysis are chiral synthetic catalysts such as transition metal complexes [1,3] and biocatalysts, specifically enzymes [4]. A significant number of industrial enantioselective processes based on enzyme catalysis are in operation [5]. Moreover, in the 1990s two important developments resulted in even greater industrial interest in the use of enzymes in asymmetric catalysis, namely, directed evolution of enantioselective enzymes [6,7] and metagenome DNA panning [8,9] (Figure 9.1).

Directed evolution involves the proper combination of molecular biological methods for random gene mutagenesis and gene expression [10], coupled with appropriate high-throughput screening systems [11], which allow rapid determination of the enantiomeric purity of a chiral product. Typically, thousands of samples

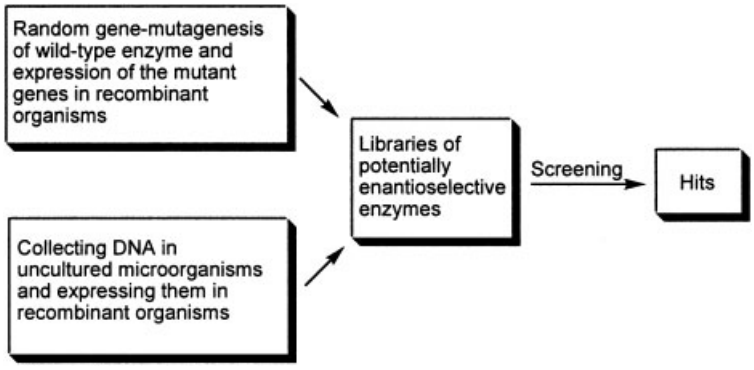

Random gene-mutagenesis of wild-type enzyme and expression of the mutant genes in recombinant organisms
Collecting DNA in uncultured microorganisms and expressing them in recombinant organisms
Libraries of potentially enantioselective enzymes
Screening
Hits

Fig. 9.1. Two sources of large libraries of potentially enantioselective enzymes.

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst
Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-30799-0

arising from the catalytic action of the evolved enzyme variants on a given substrate of interest need to be assayed within a reasonable time span, ideally one day. A similar analytical problem arises in metagenome DNA panning, in which large numbers of genes are collected in the environment, followed by expression of the encoded enzymes in recombinant microorganisms.

The enantioselectivity of a wild-type enzyme in a given transformation is traditionally determined by the so-called *ee* value of the product or, in kinetic resolution of a racemate, by the selectivity factor *E* [12]. Normally, gas chromatography (GC) or HPLC based on chiral columns is employed, but the conventional forms of these analytical tools can handle only a few dozen samples per day. Therefore, high-throughput *ee* assays had to be developed, as in the also-new field of combinatorial asymmetric transition metal catalysis [1c, 11]. In principle, the assays developed in the latter area can also be adapted to the needs of directed evolution of enantioselective enzymes, although to date this has not been put into practice. This chapter focuses on the most efficient and practical high-throughput *ee*-screening systems developed specifically to evaluate enantioselective enzymes, but a few other rapid *ee* assays not yet tested in biocatalysis are included as well. Due to space limitations, not all of the currently available *ee* assays are illustrated by detailed protocols. Many of the *ee*-screening systems are complementary, and no single assay is truly universal. For general information concerning high-throughput *ee*-screening systems, please see recent reviews [11] and Table 9.1. Selection, as opposed to screening, has not been developed to date in the directed evolution of enantioselective enzymes, although a screening system based on differential cell growth has been described

Table 9.1. High-throughput *ee* assays currently available.

Detection System/Description: Application	Reference
UV/visible: kinetic resolution of *p*-nitrophenol esters	6a
UV/visible/Quick-E-test: kinetic resolution of esters	14
UV/visible/pH indicator: kinetic resolution of esters	15
UV/visible/enzyme-coupled: kinetic resolution of acetates	16
UV/visible/enzyme-coupled: alcohols	17
UV/visible/enzyme immunoassays: alcohols	18
MS/diastereomer formation: many compounds	19
MS/labeled compounds: kinetic resolution or desymmetrization of compounds bearing enantiotopic groups	20, 33–36
NMR/flow-through cell: essentially any compound	21
FTIR/labeled compounds: kinetic resolution and desymmetrization	22
Circular dichroism: most compounds	23, 24
Fluorescence/capillary array electrophoresis: amines, alcohols, etc.	25
Fluorescence/enzyme coupled: kinetic resolution of esters	26
Fluorescence: kinetic resolution of alcohols	27
Fluorescence/DNA microarrays: amino acids, etc.	28
GC/special construction: volatile compounds	29
IR thermography: kinetic resolution in general	30, 31

[13]. Colony-based *ee* assays have not been developed so far. Rather, all the assays described in this chapter are carried out in the wells of microtiter plates following colony picking.

9.2 Materials and Equipment

9.2.1 Assays Based on Mass Spectrometry

9.2.1.1 Directed Evolution of a Lipase for Desymmetrization of *meso*-1,4-Diacetoxycyclopentene [33]

- Oligonucleotide primers for amplification of the wild-type lipase gene of *Bacillus subtilis*
- Suitable expression vector for expression in *E. coli*
- Enzymes: *Taq* DNA polymerase, restriction endonucleases, T4 DNA ligase
- LB/M9 media
- Antibiotics, additives: carbenicillin, isopropyl-β-D-thiogalactoside (IPTG)
- Organic solvents: ethanol (100%), methanol, DMSO
- Phosphate buffer: 10 mM, pH 7.5
- Sodium acetate: 10 mM
- pseudo meso compound **1** (see Section 9.2.2)
- *meso*-1,4-diacetoxycyclopentene (Fluka, Buchs, Switzerland)
- Deep-well microtiter plates (glass and plastic)

9.2.1.2 Synthesis of Pseudo meso Compound 1 [20]

- (1*S*,4*R*)-*cis*-4-acetoxy-2-cyclopenten-1-ol (**2**; prepared according to a known procedure [20])
- Organic solvents: pyridine, dichloromethane, hexane, ethyl acetate
- D_3-acetyl chloride
- Extraction: HCl (1 M), $NaHCO_3$ (saturated), NaCl (saturated), $MgSO_4$
- Chromatography: silica gel

9.2.1.3 Automation Required for High-throughput *ee* Screening

- Deep-well plates
- LB media
- Robot, e. g., Colony Picker Q-Pix (Genetix, New Milton, UK)
- 8-channel dispenser, e. g., Dispenser Multidrop DW (Thermo Electron, Vantaa, Finland)
- Pipetting robot, e. g., Genesis, and Gemini and Facts software (Tecan, Maennedorf, Switzerland)

- Software Masslynx 3.5, Quanlynx, Openlynx, and Openlynx Browser (Micromass, Manchester, UK)
- Multiplexed sprayer system (Micromass)

9.2.2 Assays Based on NMR Spectrometry

9.2.2.1 Implementation of High-throughput NMR Assay

- (*S*)-^{13}C-1-phenylethyl acetate ((*S*)-^{13}C-**4**) and (*R*)-1-phenylethyl acetate-**4** ((*R*)-**4**) (prepared according to a known procedure [21])
- Deuterated solvent, e. g., $CDCl_3$, D_6-DMSO, or D_2O
- Flow-through NMR cells for, e. g., BEST^{TM} (Bruker Biospin GmbH, Rheinstetten, Germany) or VAST^{TM} (Varian, Palo Alto, CA, USA) spectrometer (300 MHz)
- Autosampler, e. g. Gilson 215 (Gilson, Middleton, WI, USA)
- Software AMIX^{TM} (Bruker Biospin)

9.2.3 Assay Based on FTIR Spectroscopy

9.2.3.1 Determination of Molar Coefficients of Absorbance of Labeled and Unlabeled 1-Phenylethyl Acetates [22]

- (*R*)-1-phenylethyl acetate ((*R*)-**4**), and (*S*)-1-phenylethyl acetate ((*S*)-**4**) (prepared according to a known procedure [21])
- Cyclohexane
- FTIR spectrometer (Bruker Optik GmbH, Ettlingen, Germany)

9.2.3.2 High-throughput *ee* Determination by FTIR Assay

- HTS-FTIR system, e. g., Tensor 27 FTIR spectrometer, coupled to the HTS-XT system, controlled by the software OPUS© and OPUS Lab© (Bruker Optik)
- Microplate stacking device, e. g., Twister 1 (Zymark, Hopkinton, MA, USA)
- Autosampler, e. g., Microlab 4000 (Hamilton, Bonaduz, Switzerland)

9.2.4 Assays Based on UV/Visible Spectroscopy

- *N,N*-bis(2-hydroxyethyl)-2-(aminoethanesulfonic acid) (BES; Sigma, Steinheim, Germany)
- *p*-nitrophenol (Fluka)
- (*R*)-solketal butyrate, and (*S*)-solketal butyrate ((D)+(L)-solketal, Fluka, prepared according to a known procedure [21])
- Acetonitrile (Merck, Darmstadt, Germany)
- UV/visible plate reader (Spectramax, Molecular Devices, Sunnyvale, CA, USA)

9.2.5 Enzyme-coupled UV/Visible-based Assay for Hydrolases

- Test kit for the determination of acetic acid (R-Biopharm GmbH, Darmstadt, Germany)
- *Candida antarctica* lipase B (CAL-B) (Chirazyme L2, Roche, Basel, Switzerland)
- Sodium phosphate buffer: 10 mM, pH 7.3
- *Pseudomonas fluorescens* lipase AK (Amano Pharmaceutical Co. Ltd., Nagoya, Japan)
- UV/visible photometer, e. g., Ultrospec 3000 (Pharmacia Biotech Ltd., Uppsala, Sweden).
- Optional: microplate fluorescence reader, e. g., FLUOstar (BMG LabTechnologies, Offenburg, Germany)

9.3 Protocols

9.3.1 Assays Based on Mass Spectrometry

Enantiomers have identical mass spectra, which means that the relative amounts of the (*R*) and (*S*) forms present in a given sample and therefore the *ee* value cannot be measured by conventional mass spectrometry (MS) techniques. However, two MS-based approaches have been described, which allow *ee* determination. In the first method [19] two conditions have to be met: (1) a mass-tagged chiral derivatization agent is applied to the mixture, and (2) a significant degree of kinetic resolution occurs during derivatization (Horeau's principle [32]). The relative amounts of mass-tagged diastereomers can then be measured by MS simply by integrating the appropriate peaks, the uncertainty in the *ee* value amounting to $\pm 10\%$ [19]. High-throughput application (e. g., in enzyme catalysis) has not been demonstrated, but it should be possible.

The second MS-based approach does not require any derivatization reaction and has in fact been applied several times in the area of directed evolution [20, 33–36]. It makes use of deuterium-labeled pseudo enantiomers or pseudo meso compounds. This practical method is restricted to studies involving kinetic resolution of racemates and desymmetrization of prochiral compounds bearing reactive enantiotopic groups (Figure 9.2) [20].

The products of these transformations are pseudo enantiomers differing in absolute configuration *and* in mass, integration of the MS peaks and data processing affording the *ee* or *E* values. Any type of ionization can be employed, but electrospray ionization (ESI) is used most commonly [20, 33–35]. An internal standard is advisable if it is necessary to determine percent conversion. The uncertainty in the *ee* value is less than $\pm 5\%$. In the original version about 1000 *ee* values could be measured per day [20a], but this has recently been increased to about 10 000 sam-

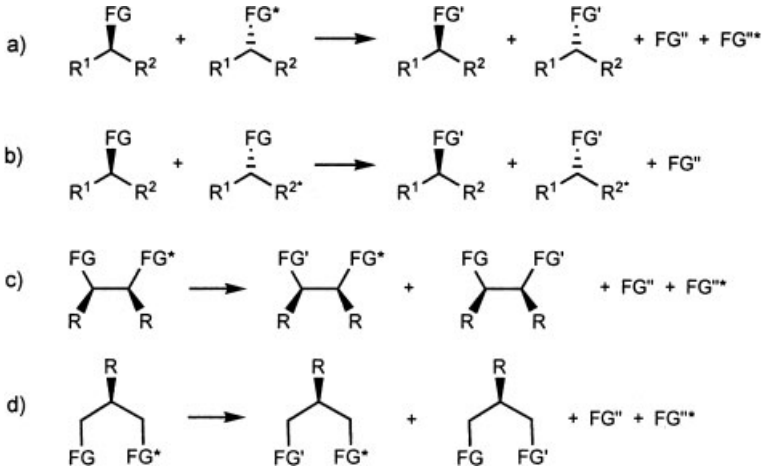


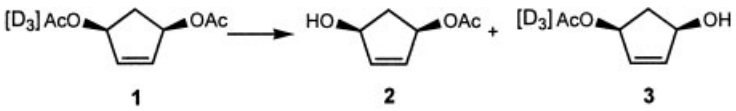
Fig. 9.2. (a) Asymmetric transformation of a mixture of pseudo enantiomers involving cleavage of the functional groups FG and labeled FG*. (b) Asymmetric transformation of a mixture of pseudo enantiomers involving either cleavage or bond formation at the functional group FG; isotopic labeling at R^2 is indicated by the asterisk. (c) Asymmetric transformation of a pseudo meso substrate involving cleavage of the functional groups FG and labeled FG*. (d) Asymmetric transformation of a pseudo prochiral substrate involving cleavage of the functional groups FG and labeled FG* [20].

ples per day with a second-generation system based on an 8-channel multiplexed sprayer system [20b]. The method is illustrated here using lipase variants from *Bacillus subtilis,* produced by the methods of directed evolution and employed as catalysts in desymmetrization of *meso*-1,4-diacetoxycyclopentene. The goal is to obtain enantioselective variants of this lipase which are expressed in *E. coli* [33]. This particular substrate and the MS-based *ee* assay have also been used in another study concerning assembly of designed oligonucleotides (ADO) as a new recombinant method in directed evolution [34]. Accordingly, the D_3-labeled pseudo meso compound **1** is used as the substrate, the two products of the asymmetric transformation being nonlabeled, and the D_3-labeled pseudo enantiomers **2** and **3** being easily distinguished by ESI-MS. It is instructive to describe all the steps, starting from gene expression and ending in *ee* determination and identification of amino acid substitutions of enhanced variants.

The MS assay has also been applied successfully in the directed evolution of enantioselective epoxide hydrolases acting as catalysts in the kinetic resolution of chiral epoxides [35]. Moreover, Diversa has recently employed the MS-based technique for desymmetrization of a prochiral dinitrile catalyzed by mutant nitrilases [36]. In this industrial application one of the nitrile moieties was labeled with ^{15}N, which means that the two pseudo enantiomeric products differ by only one mass unit.

It should be noted that in kinetic resolution, the MS measurements must be performed in the appropriate time window (near 50% conversion). If this is difficult

to achieve due to different activities of the mutants being screened, the system needs to be adapted in terms of time resolution. This means that samples for MS evaluation need to be taken as a function of time. Finally, it is useful to point out the possibility of multi-substrate *ee* screening, which allows for enzyme fingerprinting with respect to the enantioselectivity of several substrates simultaneously.



Scheme 1

9.3.1.1 Protocol for Individual Steps in Directed Evolution of a Lipase for Desymmetrization of *meso*-1,4-Diacetoxycyclopentene [33]

1. Apply a standard error-prone PCR (epPCR; see Chapter 2) to the wild-type lipase gene from *Bacillus subtilis* and express conventionally in *E. coli* [37]; initiate by inoculation of the cultures in deep-well microtiter plates (96-well format). Use LB/M9 medium with 100 µL carbenicillin (100 mg mL^{-1}) per 100 mL of medium and incubate for 5–6 h at 37 °C while shaking.
2. Induce with isopropyl-β-D-thiogalactoside (IPTG solution): 10 µL of a sterile filtered solution (100 mg mL^{-1}) in pure ethanol. Incubate overnight at 37 °C.
3. Centrifuge the cultures (10 min at 4000 rpm).
4. Prepare reaction solutions of 125 µL phosphate buffer (10 mM; pH 7.5), 50 µL supernatant, and 25 µL substrate (**1**) solution (0.1 M in DMSO; see Section 9.3.1.2)
5. Allow to react in the wells of 96-format microplates at room temperature in an incubator (shaking) for 24 h.
6. Extract with ethyl acetate (200 µL reaction solution + 200 µL solvent), place in deep-well glass microtiter plates, and dilute with standard solution for MS measurements (methanol/10 mM NaOAc, 4:1 plus undeuterated *meso*-1,4-diacetoxycyclopentene as internal standard).
7. Perform ESI-MS analysis [20].

9.3.1.2 Synthesis of Pseudo meso Compound 1 [20]

The stirred mixture of (1*S*,4*R*)-*cis*-4-acetoxy-2-cyclopenten-1-ol (**2**) (5.0 g; 35 mmol) and pyridine (4.18 g, 6.95 mmol) in dichloromethane (100 mL) is treated at 0 °C with commercially available D3-acetyl chloride (3.44 g; 42 mmol). The mixture is allowed to reach room temperature within 12 h and is then treated twice with 1 M HCl (50 mL), saturated aqueous $NaHCO_3$ and NaCl. The organic phase is dried over $MgSO_4$, the solvent is removed, and the crude product is purified by

flash chromatography over silica gel (hexane/ethyl acetate 5:1) to yield the product **1** (6.38 g; 97% as shown by NMR analysis).

9.3.1.3 Protocol for Automation for High-throughput *ee* Screening [20, 34]

After plating the bacterial colonies on agar plates, they are picked and placed individually in the deep wells of microtiter plates (96-format) containing LB medium with the aid of an appropriate robot. Up to 10 000 colonies can be handled per day. The LB medium is added with an 8-channel dispenser.

Preparation of the reaction solutions in the deep wells (2.2 mL) of microtiter plates (96-format) is automated by using a pipette robot. Pipette scripts (Gemini software) are used for robotically filling the wells with buffer and substrate solutions (see Section 9.3.1.1). To activate all the modules of the robot, Facts software is used. The pipette robot consists of a workstation with spaces for 12 microtiter plates, a robot arm for transport, a carrousel for storing the reaction plates, and a 96-fold pipette module (Figure 9.3).

Following lipase-catalyzed desymmetrization reactions of the substrate (e. g., **1**) on the microtiter plates, an extraction step is necessary prior to MS analysis. This process is controlled by the Facts software (Figure 9.4. Four modules are controlled simultaneously: the robot arm (RoMa), the carousel for storing the microtiter plates, the 96-pipette system (TeMo), and the 8-fold pipette head (Gemini). Iteration occurs within 12 min.

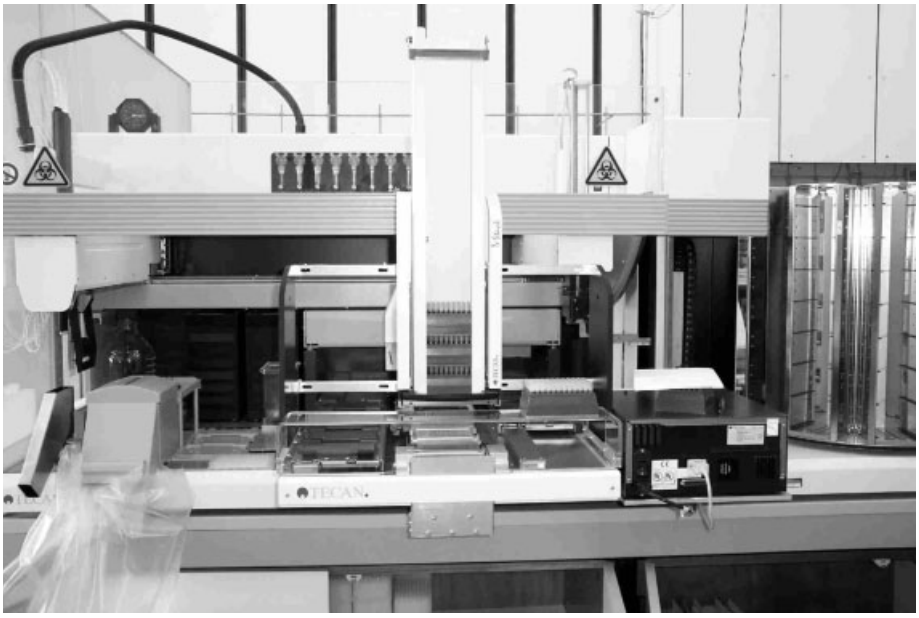


Fig. 9.3. Pipette robot Genesis (Tecan) with integrated carousel (right).

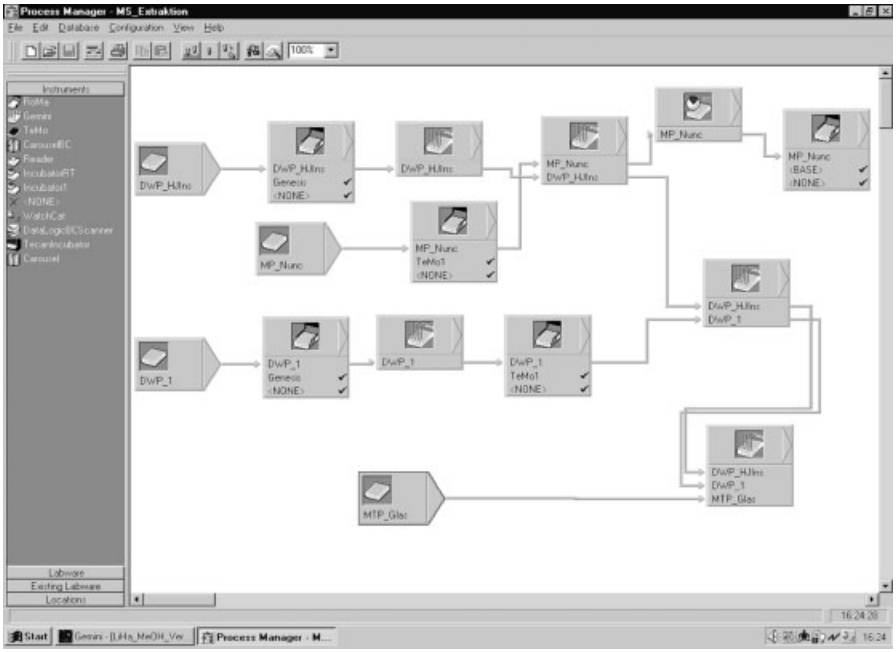


Fig. 9.4. Control of individual modules by the Facts software (Tecan). To optimize the sequence of events, the processes are time-interlocked.

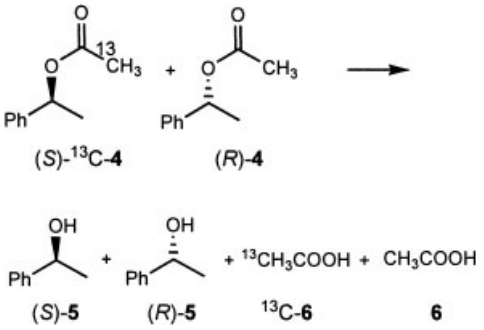
Control of the HPLC pump, the autosampler, and the MS is ensured by Masslynx 3.5 software. After optimization of the measurement conditions, a list of process measurements is set up (sample list), and the desired HPLC and MS steps are called upon. After a measurement, the ESI source is automatically brought to room temperature (shut down). Using 96-microtiter plates, 576 samples can be processed per measurement. The chromatograms are integrated by the software packages Quanlynx and Openlynx and exported as an Excel table. A macro is used to calculate the absolute intensities and therefore the *ee* and the conversion. The *E* values in kinetic resolution are automatically calculated with the formula of Sih [12]. Data processing is done with the Openlynx Browser. The overall process occurs continuously and enables analysis of up to 10 000 samples per day, provided that the 8-channel multiplexed sprayer system is used [20b]. It is also possible to use 384-well microtiter plates. Systematic optimization is required for each new compound.

9.3.2 Assays Based on NMR Spectroscopy

Traditionally, NMR measurements are considered to be slow processes, but recent advances in the design of flow-through cells have allowed the method to be applied in combinatorial chemistry [38]. These technological improvements were then ap-

plied to the development of two different NMR-based high-throughput *ee* assays [21]. In one version, classical derivatization with a chiral reagent or NMR-shift agent is parallelized, about 1400 *ee* measurements being possible per day with a precision of ±5%. In the second version, illustrated here in detail, a principle related to that of the MS system described in Section 9.3.1 is applied, i. e., chiral or meso substrates are labeled so as to produce pseudo enantiomers or pseudo meso compounds, which are then used in the actual screen. Application is thus restricted to kinetic resolution of racemates and desymmetrization of prochiral compounds bearing reactive enantiotopic groups (Figure 9.2).

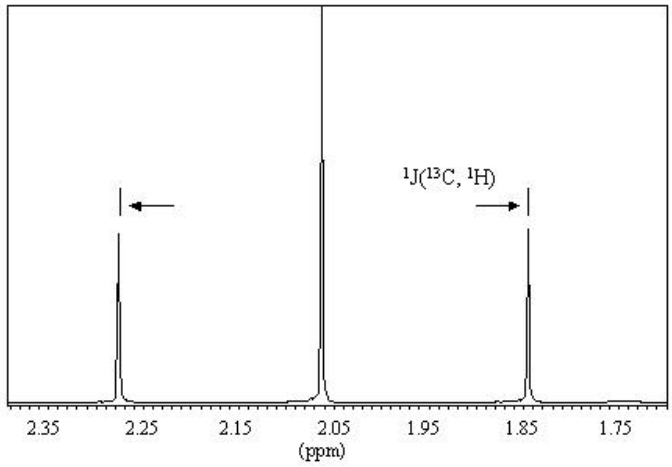
A particularly practical form of this assay utilizes ^{1}H NMR spectroscopy, ^{13}C labeling being used to distinguish between the (*R*) and (*S*) forms of a chiral compound. Essentially any carbon atom in the compound of interest can be labeled (Figure 9.2), but methyl groups in which the ^{1}H signals are not split by ^{1}H,^{1}H coupling are preferred, because the relevant peaks to be integrated are the singlet arising from the CH_3 group of one enantiomer and the doublet of the $^{13}CH_3$ group of the other. A typical example, which illustrates the method, concerns the lipase- or esterase-catalyzed hydrolytic kinetic resolution of *rac*-1-phenylethyl acetate, derived from *rac*-1-phenylethanol. However, the acetate of any chiral alcohol or the acetamide of any chiral amine can be used. Labeling can be carried out at any position of a compound, as in (*S*)-^{13}C-**4**. The synthesis is straightforward, since it simply involves acylation of the (*S*)-alcohol using commercially available ^{13}C-labeled acetyl chloride. Then a 1:1 mixture of labeled and unlabeled compounds (*S*)-^{13}C-**4** and (*R*)-**4** is prepared, which simulates a racemate. It is used in the actual catalytic hydrolytic kinetic resolution, which affords a mixture of true enantiomers (*S*)-**5** and (*R*)-**5**, as well as labeled and unlabeled acetic acid, ^{13}C-**6** and **6**, respectively, together with unreacted starting esters. At 50% conversion (or at any other point of the reaction) the ratio of (*S*)-^{13}C-**4** to (*R*)-**4** reveals the enantiomeric purity of the unreacted ester, and the ratio of ^{13}C-**6** to **6** correlates with the relative amounts of (*S*)-**5** and (*R*)-**5**.



Scheme 2

Figure 9.5a shows an excerpt of the ^{1}H NMR spectrum of a 'racemic' mixture of (*S*)-^{13}C-**4** and (*R*)-**4**, featuring the expected doublet of the ^{13}C-labeled methyl group and the singlet of the nonlabeled methyl group. Figure 9.5b displays the singlet of the nonlabeled methyl group of (*R*)-**4**, including the ^{13}C satellites due to the presence of natural ^{13}C in the sample [21].

a)



b)

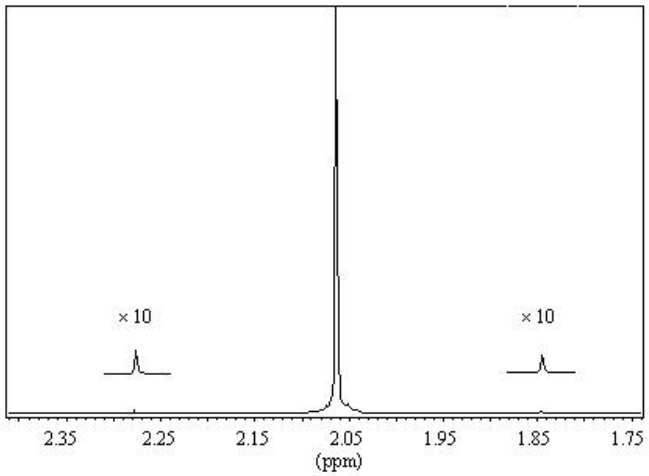


Fig. 9.5. Expanded region of the ^{1}H NMR spectra of (a) racemic mixture of (*S*)-^{13}C-**4**/(*R*)-**4** and (b) (*R*)-**4** alone [21].

The exact ratio of the two pseudo enantiomers is accessible by simple integration of the respective peaks, which provides the *ee* value. Quantitative analysis can be accomplished automatically by suitable software such as AMIX^{TM} (Bruker Biospin). The presence of naturally occurring ^{13}C in the nonlabeled (*R*) substrate is automatically considered in the dataprocessing step. As demonstrated in control experiments, the agreement with the corresponding *ee* values obtained by independent GC analysis is excellent, the correlation coefficient amounting to 0.9998 [21].

Because the *ee* value in an actual kinetic resolution depends on the degree of conversion, the selectivity factor *E* needs to be ascertained, which is possible if the conversion can be measured (see Section 9.3.1). In the present system this can be accomplished by automatic integration of the corresponding methine signals of the unreacted substrate ester at 5.9 ppm and the product alcohol at 4.9 ppm. Then the *E* value can be estimated according to the method of Sih [12]. In other cases an internal standard may be more appropriate. The precision in the *ee* values amounts to ±2% as checked by independent GC analysis. For the first version of a high-throughput *ee* assay based on traditional derivatization with chiral reagents such as Mosher's acid chloride, the same equipment and software can be used. Again, about 1400 samples can be handled per day, precision in the *ee* value being ±5% [21]. Thus, these two NMR-based *ee*-screening systems are practical, precise, and rather general.

9.3.2.1 Protocol for Implementing High-throughput NMR Assay [21]

The enzymatic reactions are performed in the wells of microtiter plates (96-format) in water (as in lipase-catalyzed hydrolytic reaction of (*S*)-^{13}C-**4**/(*R*)-**4**), which is followed by a standard automatic extraction step. Depending on the particular substrate to be assayed and the type of solvent used, it may be necessary to remove the solvent. However, this is often not necessary. For enzymatic reactions in organic medium, solvent extraction is not required. For NMR analysis such solvents as $CDCl_3$, D_6-DMSO, or D_2O are used. A minimum of about 6 µmol of substrate/product per milliliter of solvent is needed. Although the flow-through cell system does not need too much solvent (about 1 L in 24 h), the solvents can be mixed with the undeuterated form in 1:9 ratios to reduce costs.

Several flow-through NMR cells are commercially available, for example, the BEST^{TM}-NMR system (Bruker Biospin) described here or the VAST^{TM} NMR system (Varian) (Figure 9.6). In addition to the flow-through cell and the NMR spectrometer (300 MHz), the system requires an autosampler, for example, a Gilson 215 autosampler [21].

Here, hydrolytic kinetic resolution of the acetate of racemic 1-phenylethanol using a 1:1 mixture of the pseudo enantiomers (*S*)-^{13}C-**4** and (*R*)-**4** in water is carried out in the wells of microtiter plates (e. g., 96-format), followed by extraction (pipette robot) using 300 µL of $CDCl_3$. For storage, the resulting organic layers are placed in the wells of microtiter plates. The samples are then transferred to the autosampler of the BEST^{TM}-NMR system and analyzed using the high-speed mode as described below. The samples are taken by the movable needle and transferred into the first valve system. At the same time, washing solution ($CDCl_3$) is introduced by the Dilutor 402 into the 6-port selection valve. Injection occurs via the injection port, whereby the washing solution is first fed into the second 6-port selection valve, followed by the sample to be measured. The washing solution is then transferred rapidly via tubing into the flow-through cell by a hydraulic impulse. Immediately thereafter, the sample follows, which is separated from the washing solution by a

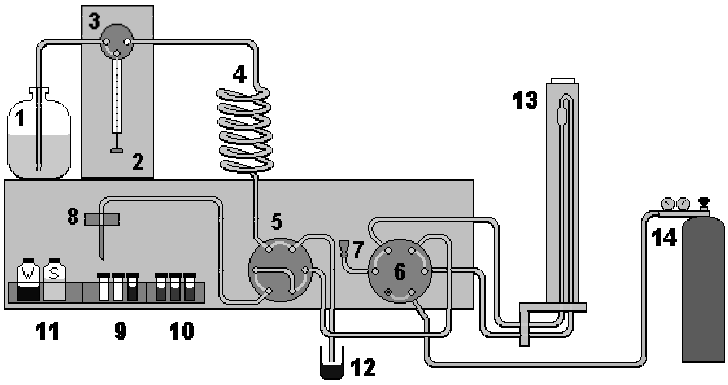


Fig. 9.6. Schematic representation of the BEST^{TM} system (Bruker Biospin; see also [21]). 1, Bottle with transport liquid; 2, dilutor 402 single syringe (5 mL) with 1100 µL tube; 3, dilutor 402 3-way valve; 4, sample loop (250–500 µL); 5, 6-way valve (standard version) loading sample; 6, 6-way valve (standard version) injecting sample; 7, injection port; 8, XYZ needle; 9, rack for sample vials; 10, rack for recovering vials; 11, rack for washing fluids and waste bottle (3 glass bottles); 12, external waste bottle; 13, flow probe with inner lock container; 14, inert gas pressure canister for drying process.

small air gap. The washing solution is pumped through the flow-through cell, and once the sample has entered it, pumping is stopped and the NMR measurement is automatically initiated (maximum of 4 scans). During this time the washing solution is stored behind the cell.

Because the same solvent is used for all samples, NMR locking and shimming is principally necessary only once, at the beginning of the process. However, since the shim may not be constant, locking and shimming should be repeated after about every 10th sample. Following the NMR measurement, the sample and the washing solution are flushed out of the system by automatic pumping in the reverse direction. During the NMR measurement of one sample, the next one is prepared by the autosampler. About 1400 samples can be handled per day. The spectral data is analyzed with the aid of appropriate software, for example, Software AMIX^{TM} (Bruker Biospin). For this purpose, the region of the spectrum to be integrated needs to be defined. The data that is accumulated as a result of appropriate NMR peak integration is transferred to Excel spreadsheets. The *ee* or *E* values are readily tabulated with the help of a macro.

9.3.3 Assay Based on FTIR Spectroscopy

The concept of isotopic labeling for distinguishing pseudo enantiomers in the kinetic resolution of chiral compounds and in the desymmetrization of prochiral substrates bearing reactive enantiotopic groups (Sections 9.2 and 9.3) can also be applied when Fourier transform infrared spectroscopy (FTIR) is used as the detec-

tion system [22]. Because FTIR spectroscopy is an inexpensive analytical technique available to almost all laboratories, the *ee* assay has great potential. Moreover, it is of particular interest in the analysis of enzymatic reactions, because the *ee* or *E* values can be measured directly in culture supernatants without time-consuming workup procedures. Of course, it is restricted to substrates that contain IR-active functional groups (Figure 9.4). If all prerequisites are met, up to 10 000 *ee* values can be measured per day with an accuracy of 7%, making this a particularly practical and inexpensive method for evaluating large libraries of potentially enantioselective enzymes [22].

To assess the applicability of FTIR spectroscopy for determining *ee* values for a given substrate, especially with regard to accuracy, the 'best' position at which isotopes are introduced needs to be determined. For illustration, the lipase- or esterase-catalyzed kinetic resolution of esters is considered here, although the method is not restricted to this type of reaction. ^{13}C labeling of carbonyl groups is ideal for several reasons:

- Carbonyl groups provide intense vibrational bands in an IR spectrum, allowing for easy and precise determination of the concentration of the compounds by applying the Lambert-Beer law.
- In the spectral region between 1600 and 1800 cm^{-1}, which is typical of carbonyl-stretching vibrations, almost no absorption bands of other functional groups appear, eliminating interference.
- ^{13}C-labeled compounds can be easily prepared because reactive reagents with ^{13}C-labeled carbonyl groups like 1-^{13}C-acetyl chloride are commercially available.
- The absorption maxima of the carbonyl-stretching vibration is shifted by 40 to 50 cm^{-1} to lower wave numbers by introducing a ^{13}C label, which prevents overlap of the two carbonyl bands.

A specific example concerns the kinetic resolution of 1-phenylethyl acetate, previously used to illustrate the NMR-based *ee* assay (see Section 9.3.2). The optimal way to proceed is to apply ^{13}C labeling in the carbonyl moiety, i. e., to prepare a pseudo racemate comprising a 1:1 mixture of (*S*)-^{13}C-**4** and (*R*)-**4** (Section 9.3). Figure 9.7 shows part of the FTIR spectrum of a 1:1 mixture of (*R*)-**4** and (*S*)-^{13}C-**4**, illustrating the anticipated shift of the respective carbonyl-stretching vibration, which allows quantification of the pseudo enantiomers [22].

To apply the Lambert–Beer law in calculation of the concentrations of the pseudo-enantiomeric substances, the molar coefficients of absorbance need to be determined. For this purpose, solutions of (*R*)-**4** and (*S*)-^{13}C-**4** in cyclohexane at different concentrations have to be prepared. After measuring the corresponding absorbances at the absorption maxima of the carbonyl-stretching vibration, the molar coefficients of absorbance are calculated by applying the Lambert-Beer law: $E = \varepsilon \cdot c \cdot d$ (Figure 9.8).

With these coefficients in hand, exploitation of the FTIR spectra of different synthetic mixtures of the labeled and unlabeled enantiomeric compounds is possible. After applying an automated baseline correction to the spectra and correcting the

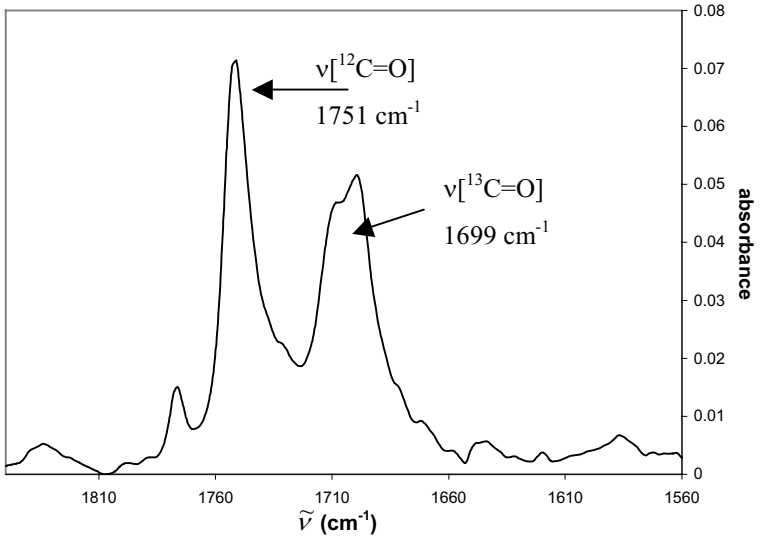


Fig. 9.7. Part of an FTIR spectrum of a 1:1 mixture of (*R*)-**4** and (*S*)-^{13}C-**4** [22].

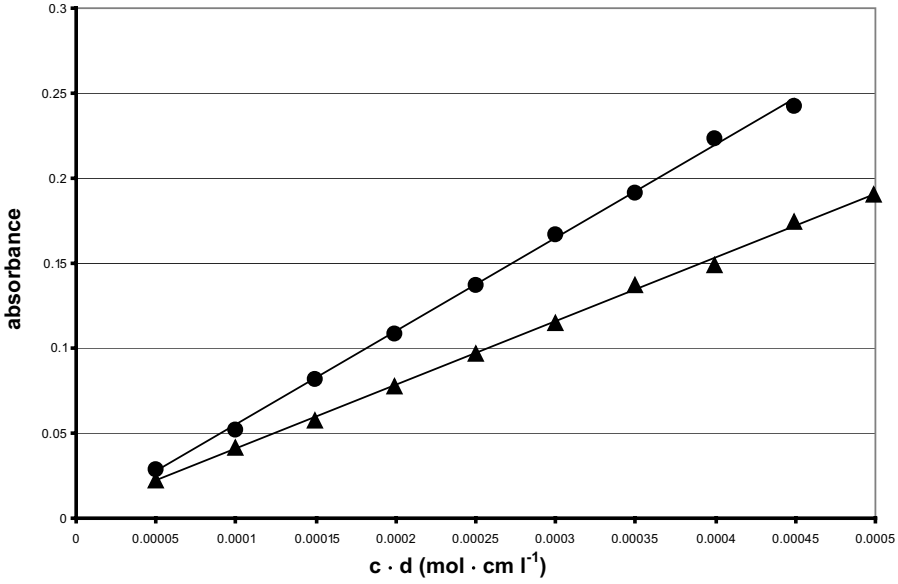


Fig. 9.8. Determination of the molar coefficients of absorbance of (*R*)-**4** (circles, absorption maximum: 1751 cm^{-1}) and (*S*)-^{13}C-**4** (triangles, absorption maximum: 1699 cm^{-1}) by linear regression [22].

absorbance of one enantiomer in the synthetic mixtures by the absorbance of the other enantiomer at this position, the accuracy of the pseudo-enantiomeric system based on 1-phenylethyl acetate is excellent, specifically within ±3% in comparison to the *ee* values determined by chiral GC [22].

High-throughput measurements are possible with commercially available HTS–FTIR systems. The analysis can be performed in a Tensor 27 FTIR spectrometer coupled to a HTS–XT system that can analyze samples on 96- or 384-well microtiter plates. The plates are equipped with a silicon plate for IR transmittance. Moreover, the *ee* values can be measured in culture supernatants, which is not possible with MS- or NMR-based assays (Sections 9.3.1 and 9.3.2).

For this purpose, Bruker has already coupled the microplate stacking device Twister 1 to its microplate reader [22]. In this combination, which is controlled by OPUS© software, 40 IR microplates can be measured automatically. To load samples with high throughput, the Microlab 4000 SP autosampler can be used. Both formats (96 and 384) of the Bruker silicon microplates are suitable for automatic loading of various types of samples (proteins, cells, culture media).

9.3.3.1 Protocol for Determining Molar Coefficients of Absorbance of Labeled and Unlabeled 1-Phenylethyl Acetates [22]

After preparation of a stock solution (0.200 M) of (*R*)-1-phenylethyl acetate ((*R*)-**4**) and (*S*)-(1-phenylethyl)-1-^{13}C-acetate ((*S*)-^{13}C-**4**) in cyclohexane, the solutions are diluted with cyclohexane to concentrations of 0.180, 0.160, 0.140, 0.120, 0.100, 0.080, 0.060, 0.040, and 0.020 M (total volume: 1 mL). The absorbance of the resulting samples is measured with a FTIR spectrometer at the corresponding absorption maxima of the carbonyl-stretching vibration ((*R*)-**4**: 1751 cm^{-1}; (*S*)-^{13}C-**4**: 1699 cm^{-1}) with a thickness of the layers of 25.0 µm, performing 32 scans at a resolution of 4 cm^{-1}. The molar coefficients of absorbance are determined by linear regression, with correlation coefficients >0.995. Analysis of synthetic mixtures of the pseudo enantiomers of 1-phenylethyl acetate is performed under the same conditions at a concentration of 0.10 M.

9.3.3.2 Protocol for High-throughput *ee* Determination by FTIR Assay [22]

High-throughput measurements are performed in a Tensor 27 spectrometer connected to a HTS-XT system (Bruker Optik). Each supernatant mixture (3 µL) is transferred to a 384-well microtiter plate equipped with a silicon plate for IR transmittance. Every sample is measured with a resolution of 8 cm^{-1} and a scan number of 10, so the total time for the analysis of each sample is 8.9 s, allowing a throughput of >9000 samples per day. The resulting spectra are analyzed with Opus© and Opus Lab© software. The first 14 samples are used for calibration, and the remaining

probes are used as unknown samples. To evaluate the accuracy of the system, the *ee* value of each mixture is independently determined by chiral GC analysis.

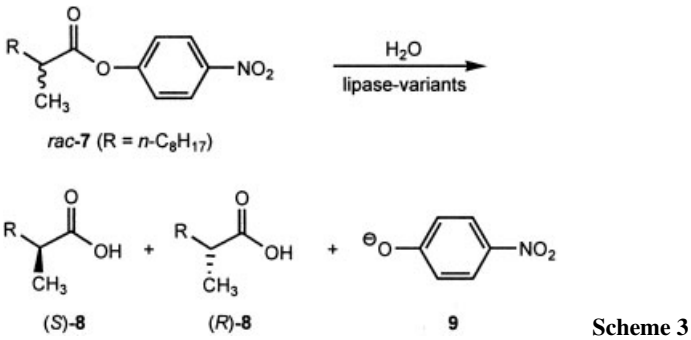
9.3.4 Assays Based on UV/Visible Spectroscopy

Color tests have several advantages, including the possibility of visual prescreening on microtiter plates. Moreover, if a reliable UV/visible signal arises in consequence of an enzymatic reaction, commercially available (and fairly inexpensive) UV/visible plate readers can be used to screen thousands of mutant enzymes.

9.3.4.1 Screening Hydrolases in Kinetic Resolution of Chiral *p*-Nitrophenol Esters

The first high-throughput *ee* assay used in the directed evolution of enantioselective enzymes was based on UV/visible spectroscopy [6a]. It was a rather crude system restricted to the hydrolytic kinetic resolution of chiral *p*-nitrophenol esters catalyzed by lipases or esterases. This assay is described here because of its simplicity and because it illustrates a principle that forms the basis of several other *ee* assays developed later. To evaluate thousands of lipase variants from *Pseudomonas aeruginosa* as potential biocatalysts in the hydrolytic kinetic resolution of chiral esters, the *p*-nitrophenol ether (*S*)-**7**/(*R*)-**7** is prepared as a model substrate. Lipase-catalyzed hydrolysis in buffered medium generates *p*-nitrophenolate (**9**), which shows a strong light absorption at 405 nm. Thus, reactions can be carried out on microtiter plates, a simple UV/visible plate reader measuring absorption as a function of time (typically during the first 8 min). However, since the racemate delivers information concerning only the overall rate, the (*S*) and (*R*) substrates are prepared and studied *separately* pairwise in 96-well microtiter plates. If the slopes of the absorption/time curves differ considerably, a hit is indicated, i. e., an enantioselective lipase variant has been identified, which is then studied in detail in a laboratory-scale reaction using traditional chiral GC. Figure 9.9 shows two experimental plots, illustrating the presence of a nonselective lipase (top) and a hit (bottom). About 500–800 plots of this kind are possible per day. By using epPCR [6a], saturation mutagenesis [6b], and DNA shuffling [6c], a total of 40 000 lipase variants were generated and screened in the model reaction [6]. Several enantioselective lipase variants were obtained, the best one showing an *E* value of >51. The wild-type lipase displays an *E* value of only 1.1. Reversal of enantioselectivity was also achieved, a process in which again about 40 000 mutants were screened using the method described [6d].

The disadvantage of this assay has to do with the fact that a built-in chromophore is required (*p*-nitrophenol), yet *p*-nitrophenol esters are never used in real (industrial) applications. Moreover, since the (*S*) and (*R*) substrate are tested separately pairwise, the enzyme does not compete for the two substrates, rendering the assay rather crude.



Scheme 3

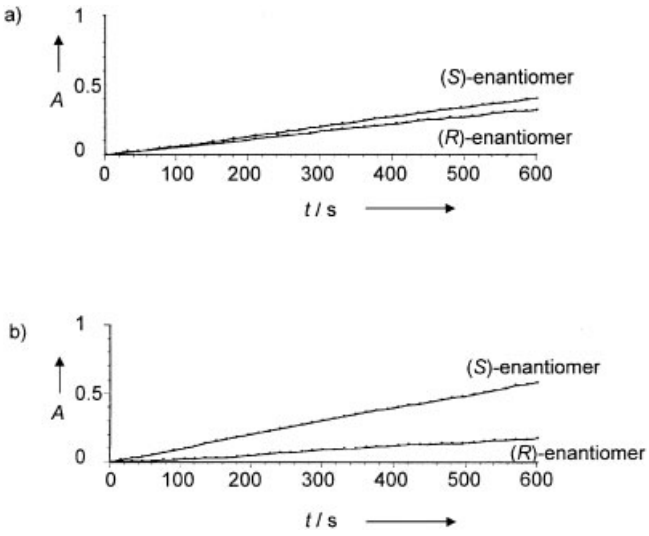
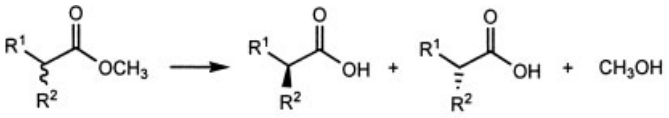


Fig. 9.9. Time course of lipase-catalyzed hydrolysis of the (*R*)- and (*S*)-ester **7** measured with a UV/visible plate reader [6a]. (a) Wild-type lipase from *P. aeruginosa*, (b) improved mutant in the first generation.

9.3.4.2 General Assay for Kinetic Resolution of Esters (Kazlauskas Test)

Because the ester hydrolysis leads to a change in acidity, as in hydrolytic lipase- or esterase-catalyzed kinetic resolution, an appropriate pH indicator can be used for quantification [14, 15]). In an optimized version (Kazlauskas test) [15], a linear correlation between the amount of acid generated and the degree of protonation of the indicator was ensured by using a buffer (e. g., *N*,*N*-bis(2-hydroxyethyl)-2-(aminoethanesulfonic acid) (= BES), and a pH indicator (e. g., *p*-nitrophenol) having the same pK_a value. The advantage of this system relates to the fact that *p*-nitrophenol esters are not necessary, i. e., 'normal' substrates such as methyl esters **10** can be used.

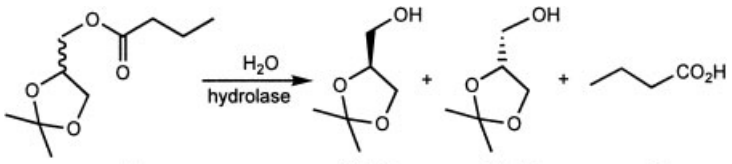


rac-**10** **11** **12** **13** **Scheme 4**

In this assay, as in related systems based on other indicators [14, 15b, 15c], the original idea of using the (*S*) and (*R*) substrates pairwise separately on microtiter plates [6a], needs to be applied. Then hits are easily identified by using a standard plate reader or even visual inspection. Dozens (or more) of microtiter plates (96-well-, or 384-well-format) can be processed per day.

9.3.4.3 Experimental Details of the Kazlauskas Test [15a]

The solutions for the assay are prepared by mixing a chiral ester, for example, (*R*)- and (*S*)-solketal butyrate (**14**), separately pairwise (420 µL of a 30.0 mM solution in acetonitrile), acetonitrile (470 µL), 4-nitrophenol (6000 µL of a 0.9115 mM solution in 5.0 mM BES, pH 7.2), and BES buffer (5110 µL of a 5.0 mM solution, pH 7.2). Hydrolase solutions (5 µL well^{-1}) are transferred from the mother plate to a 96-well microtiter plate with an 8-channel pipette. Assay solution (100 µL well^{-1}) is quickly added to each well with a 1200 µL 8-channel pipette. The final concentrations in the aqueous solutions containing 7.1% acetonitrile in each well are 1.0 mM substrate, 4.65 mM BES, and 0.434 mM 4-nitrophenol. The plate is quickly placed in the microplate reader and shaken for 10 s to ensure complete mixing, and the decrease in absorbance at 404 nm is monitored at 25 °C as often as permitted by the microplate software, typically every 11 s. The starting absorbance is typically 1.2. Data are collected for 1 h. Each hydrolysis reaction is carried out in quadruplicate and is averaged. The first 10 s of data are sometimes erratic, possibly due to dissipation of bubbles created during shaking. For this reason, the first 10 s of data are excluded from calculation of the initial rate. Activities are calculated from slopes of the linear portion of the curve, usually over the first 200 s. The initial rates are calculated from the average d*A*/d*t* [15a].



rac-**14** (*S*)-**15** (*R*)-**15** **16** **Scheme 5**

9.3.5 Enzyme-coupled UV/Visible-based Assay for Hydrolases

If the actual product of an enzymatic reaction under study can be transformed by another enzyme into a secondary product that gives rise to a spectroscopic signal, an enzyme-coupled assay is possible. This was first demonstrated using fluorescence as the spectroscopic detection method, high-throughput also being possible [26a] (see Section 9.3.6.7). Specifically, chiral esters containing a fluorogenic moiety (umbelliferone) are subjected to enzyme-catalyzed hydrolysis, and the initial product (alcohol) is then degraded enzymatically with formation of a product (umbelliferone) detectable by fluorescence. In enantioselective hydrolases, this idea, coupled with the concept of employing the (*S*) and (*R*) substrates separately pairwise, led to the establishment of a useful high-throughput *ee* assay for hydrolases [26]. The only disadvantage of this otherwise elegant approach relates to the necessity of incorporating a fluorogenic moiety in the starting material, which means that if directed evolution is performed, the final result will be specific to a substrate that may not be industrially acceptable. Instead of enzyme coupling, a chemical step leading to an absorption or fluorescence signal can also be used [26b, 26c, 26d].

In a different approach, the hydrolase-catalyzed kinetic resolution of chiral acetates was studied using a high-throughput *ee* assay also based on an enzyme-coupled test, the presence of a fluorogenic moiety not being necessary [16]. The assay is based on the idea that the acetic acid formed by hydrolysis of a chiral acetate can be transformed stoichiometrically into NADH in a series of coupled enzyme reactions using commercially available enzyme kits (Fig. 9.10). The NADH is then

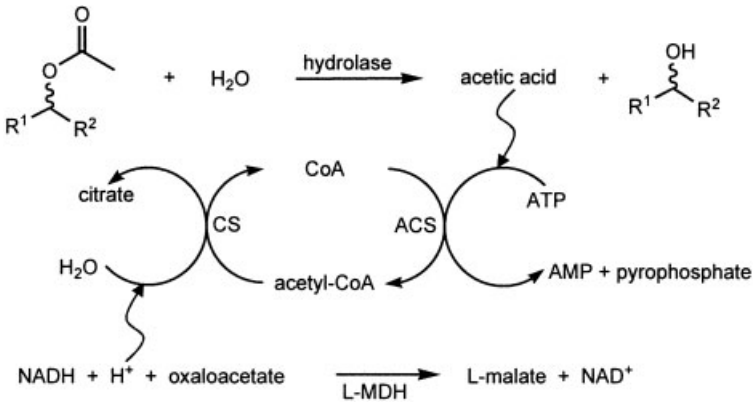


Fig. 9.10. The hydrolase-catalyzed reaction releases acetic acid, which is converted by acetyl-CoA synthetase (ACS) to acetyl-CoA in the presence of ATP and coenzyme A (CoA). Citrate synthase (CS) catalyzes the reaction between acetyl-CoA and oxaloacetate to give citrate. The oxaloacetate required for this reaction is formed from L-malate and NAD^+ in the presence of L-malate dehydrogenase (L-MDH). Initial rates of acetic acid formation can be determined by the increase in adsorption at 340 nm due to the increase in NADH concentration. Use of optically pure (*R*)- or (*S*)-acetates allows determination of the apparent enantioselectivity E_{app} [16].

easily detected by UV spectroscopy in the wells of microtiter plates with a standard plate reader. About 13 000 samples can be evaluated per day. The kinetic resolution of (*S*,*R*)-1-methoxy-2-propylacetate was studied with various commercially available hydrolases. The agreement between the apparent selectivity factor E_{app} and the actual value E_{true} determined by GC was excellent at low enantioselectivity ($E = 1.4 - 13$), but less so at higher enantioselectivity (20% variation at $E = 80$) [16].

9.3.5.1 Details of Enzyme-coupled UV/Visible-based Assay for Enantioselective Hydrolases [16]

The test kit for the determination of acetic acid released is used according to the manufacturer's protocol (see also below). Spectrophotometric determination of NADH concentration is performed at 340 nm in the milliliter scale, e. g., in an Ultrospec 3000 photometer, and on the microliter scale in a fluorimeter, e. g. FLUOstar.

General procedure for the determination of acetic acid in a MTP assay [16] are as follows. A solution of *Pseudomonas fluorescens* esterase (PFE; 20 µL, 2 mg mL^{-1}, unless stated otherwise) or *Candida antarctica* esterase (CAL-B) (2 mg mL^{-1}) is added to a mixture of the test-kit components (150 µL). The reactions are started by adding a solution of a chiral acetate (20 µL) in sodium phosphate buffer (10 mM, pH 7.3). Mixtures of the test kit with buffer or cell lysates of noninduced *E. coli* harboring the gene encoding recombinant PFE (R-Biopharm) serve as controls. In a similar manner, reaction rates are determined by using optically pure (*R*)- and (*S*)-acetates separately. For reaction with crude cell extract, PFE is produced in microtiter plates similar to the published protocol for shaker-flask cultivation [39]. However, the cultivation volume is 200 µL per well, and cells are disrupted by two freeze-thaw cycles. Finally, cell debris is removed by centrifugation, and the supernatants are used for the assay.

9.3.6 Further Assays

9.3.6.1 Enzymatic Method for Determining Enantiomeric Excess (EMDee)

A somewhat different approach to determining the enantiopurity of a sample is based on the idea that an appropriate enzyme selectively processes one enantiomer, giving rise to a UV/visible signal [17]. An example concerns determination of the enantiopurity of chiral secondary alcohols, the (*S*) enantiomer being oxidized selectively by the alcohol dehydrogenase from *Thermoanaerobium* sp. The rate of this process can be monitored by a UV/visible plate reader due to the formation of NADPH (absorption at 340 nm), which relates to the quantity of the (*S*) enantiomer present in the mixture. About 4800 *ee* determinations are possible per day, accuracy amounting to ±10%. Although the screen was not specifically developed to evaluate chiral alcohols produced by an enzymatic process, it is conceivable that this could be possible after an appropriate extraction process.

9.3.6.2 Assay Based on DNA Microarrays

Another potentially useful *ee* assay makes use of DNA microarrays [28]. This type of technology is employed to determine relative gene expression levels on a genome-wide basis as measured by the ratio of fluorescent reporters. In *ee* assays, chiral amino acids are used as model compounds. Mixtures of a racemic amino acid are first subjected to acylation at the amino function with formation of *N*-Boc-protected derivatives. The samples are then covalently attached to amine-functionalized glass slides in a spatially arrayed manner (Figure 9.11). In a second step, the uncoupled surface amino functions are acylated exhaustively. The third step involves complete deprotection to afford the free amino function of the amino acid. Finally, in a fourth step, two pseudo-enantiomeric fluorescent probes are attached to the free amino groups on the surface of the array. An appreciable degree of kinetic resolution in the process of amide coupling is required for success [28] (Horeau's principle [32]). In this example, the *ee* values are accessible by measuring the ratio of the relevant fluorescent intensities. About 8000 *ee* determinations are possible per day, precision amounting to (10% of the actual value. Although it was not explicitly demonstrated that this *ee* assay can be used to evaluate enzymes (e. g., proteases), this should in fact be possible. The question of whether other types of substrates (and enzymes) are amenable to this type of screening also needs to be addressed.

9.3.6.3 Enzyme Immunoassays to Measure Enantiomeric Excess

High-throughput screening of enantioselective catalysts is also possible by enzyme immunoassays [18], a technology that is routinely applied in biology and medicine for other purposes. As in some of the other screening systems, this new assay was not developed specifically for enzyme-catalyzed processes. In fact, it was illustrated by analyzing (*R*)/(*S*) mixtures of mandelic acid generated by enantioselective Ru-catalyzed hydrogenation of benzoyl formic acid (**17**) (Figure 9.12). Using an antibody that binds both enantiomers enabled measuring the concentration of the reaction product, thereby allowing the yield to be calculated. The use of an (*S*)-specific antibody then makes determination of the *ee* possible (Figure 9.12).

9.3.6.4 Assays Based on Gas Chromatography or HPLC

Conventional gas chromatography (GC) based on the use of chiral stationary phases can handle only a few dozen *ee* determinations per day. In some instances GC can be modified so that, in optimal situations, about 700 exact *ee* and *E* determinations are possible per day [29]. Such medium-throughput may suffice in certain applications. The example concerns the lipase-catalyzed kinetic resolution of the chiral alcohol (*R*)- and (*S*)-**18** with formation of the acylated forms (*R*)- and (*S*)-**19**. Thousands of mutants of the lipase from *Pseudomonas aeruginosa* were created by error-prone PCR for use as catalysts in the model reaction and were then screened for enantioselectivity [29].

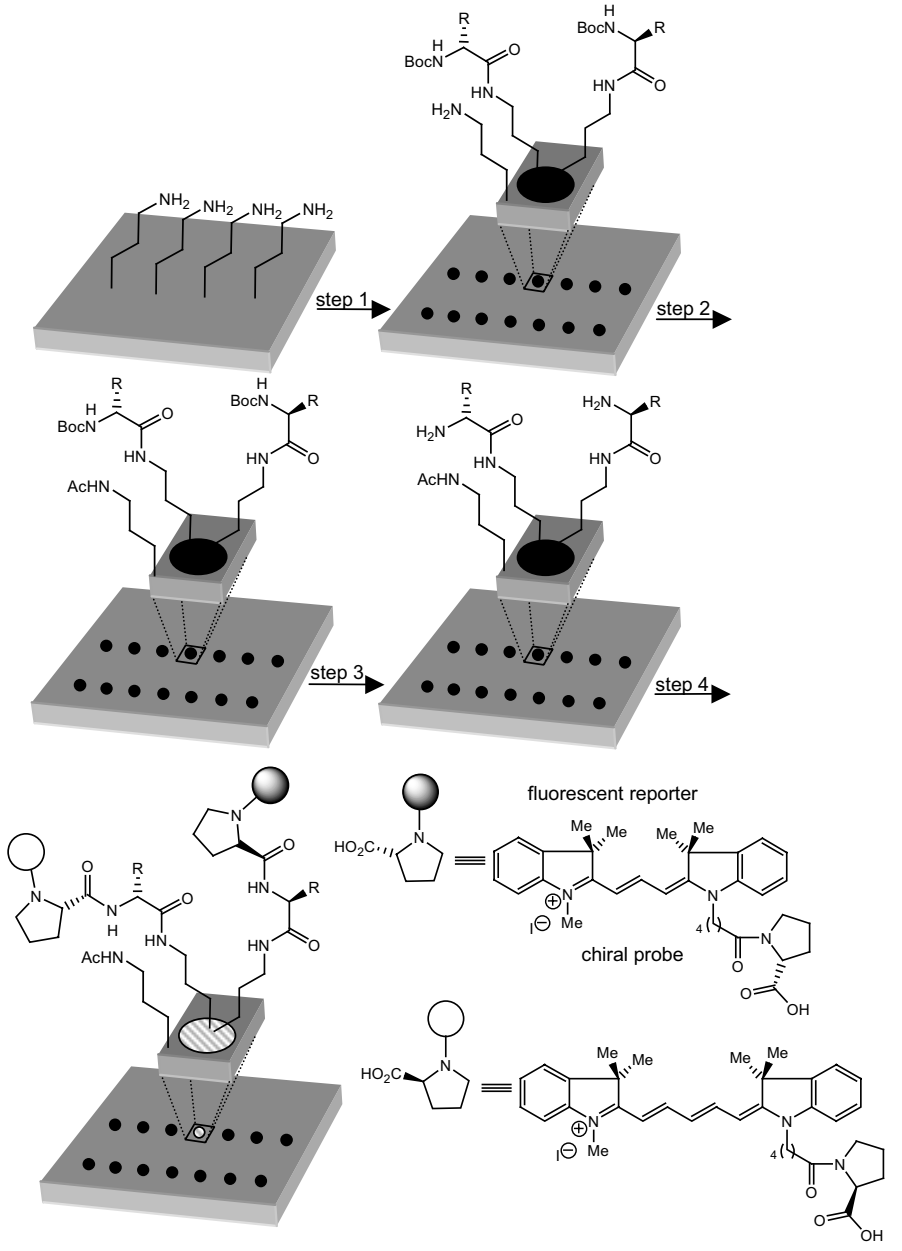


Fig. 9.11. Reaction microarrays in high-throughput *ee* determination [28]. Reagents and conditions: step 1, BocHNCH(R)CO_2H, PyAOP, *i*Pr_2NEt, DMF; step 2, Ac_2O, pyridine; step 3, 10% CF_3CO_2H and 10% Et_3SiH in CH_2Cl_2, then 3% Et_3N in CH_2Cl_2; step 4, pentafluorophenyl diphenylphosphinate, *i*Pr_2NEt, 1:1 mixture of the two fluorescent proline derivatives, DMF, −20 °C.

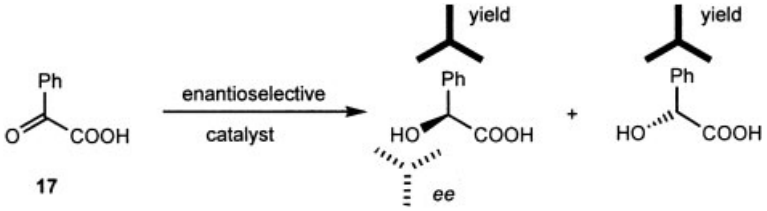
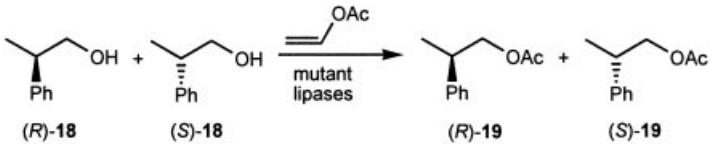


Fig. 9.12. High-throughput screening of enantioselective catalysts by competitive enzyme immunoassays [18]. The solid antibody recognizes both enantiomers, and the hatched antibody is (*S*)-specific, enabling the determination of yield and ee.



Scheme 6

The initial approach concerns the use of two columns in a single GC oven [29]. The successful setup consists of two GC instruments (e. g., GC instruments and data bus (HP-IB) commercially available from Hewlett-Packard, Waldbronn, Germany), one prep-and-load sample manager (PAL) (commercially available from CTC, Schlieren, Switzerland), and a PC. The instruments are connected to the PC via a standardized data bus (HP-IB), which controls pressure, temperature, etc., and also handles other data, such as that from the detector. A wash station and a drawer system with a maximum of 8 microtiter plates are included. The sample manager is attached to the unit so as to reach both injection ports. Because the sample manager can inject samples from 96- or 384-well microtiter plates, over 3000 samples can be handled without manual intervention. The software (Chemstation®) (Hewlett-Packard) enables additional programs (macros) to be applied before and after each analytical run. Such a macro controls the sample manager, each position on the microtiter plate is labeled via the sequence table. Another macro ensures analysis following each sample run in a specified manner, i. e., the peaks of the chiral compound **18** are analyzed quantitatively. The analytical data are transferred to an Excel spreadsheet via DDE (Dynamic Data Exchange; Microsoft) in table form or in microtiter format, allowing for a rapid overview. Finally, the setup includes H_2 guards, which monitor the hydrogen concentration in the ovens; at concentrations exceeding 1% (>4% H_2 is potentially explosive), the system responds and automatically switches to nitrogen as the carrier gas [29].

Using a stationary phase based on a β-cyclodextrin derivative (2,3-di-*O*-ethyl-6-*O*-*tert*-butyldimethylsilyl-β-CD), complete separation of (*R*)- and (*S*)-**18** (but not of (*R*)/(*S*)-**19**) is possible within 3.9 min. Because the configuration comprises two simultaneously operating GC units, about 700 exact *ee* determinations of (*R*)/(*S*)-**18** are possible per day. Moreover, the corresponding values for the conversion and the selectivity factor *E* (or *s*) are likewise automatically provided in microtiter-format, which means that the *ee* of (*R*)/(*S*)-**19** is also accessible. Of course, every

new substrate has to be optimized anew using commercially available chiral stationary phases [29]. It is sometimes better to simply use two or more separate GC instruments. A related system has been developed for HPLC analysis, specifically for use in the directed evolution of cyclohexanone monooxygenases as catalysts in the enantioselective oxidation of prochiral thioethers with formation of chiral sulfoxides (Reetz, et al., unpublished).

9.3.6.5 Capillary Array Electrophoresis

Capillary array electrophoresis (CAE) was widely used in the Human Genome Project and was recently adapted to the high-throughput *ee* determination of chiral amines [25]. The samples are first derivatized by reaction which fluorescein isothiocyanate, the derivatives then being analyzed by laser-induced fluorescence (LIF) following parallel separation on chirally modified (α- and β-cyclodextrin) capillaries. The 96-array system allows for at least 7000 *ee* determinations per day. Extension to other classes of chiral compounds needs to be demonstrated. An alternative and less expensive variation is based on CE on glass microchips [25], but high-throughput methods still need to be developed.

9.3.6.6 Assays Based on Circular Dichroism

An alternative to chiral HPLC, which separates the enantiomers of interest, is the use of normal columns that simply separate the starting materials from the enantiomeric products (and side products). The *ee* can then be determined by circular dichroism (CD). Several reports of high-throughput CD have appeared, although application to enzymatic reactions was not demonstrated [23, 24]. The potential advantages include low cost and excellent high-throughput using standard automation (typically 1000 *ee* determinations per day).

9.3.6.7 Assays Based on Fluorescence

The advantage of fluorescence-based assays is their high sensitivity. It is therefore perhaps surprising that few such systems have been developed for evaluating the enantioselectivity of enzyme-catalyzed reactions. Fluorescence as a detection method is used in an enzyme-coupled assay [26] (see Section 9.3.4.3) and in the capillary array electrophoresis [25] (see Section 9.3.6.5). If several substrates need to be screened simultaneously, fluorescence-based substrate arrays as enzyme fingerprinting tools can be used, although enantioselectivity still needs to be addressed [26e].

Another fluorescence-based method for assaying activity and enantioselectivity of synthetic catalysts, specifically in the acylation of chiral alcohols, was recently reported [27]. The idea is to use a molecular sensor that fluoresces upon formation of an acidic product (acetic acid). Adaptation to high-throughput evaluation of enantioselective lipases or esterases needs to be demonstrated.

9.3.6.8 IR Thermographic Analyses

Modern photovoltaic infrared cameras can detect heat in the form of IR radiation from objects. The picture obtained thereby provides a two-dimensional thermal image that is a spatial map of the temperature and emissivity distribution of all objects in the picture. The technique was used to test the activity of heterogeneous catalysts [40] and thereafter to detect enantioselective lipases on microtiter plates [30,31]. The method is useful for identifying highly enantioselective hits. However, because quantification has not yet been achieved, the assay cannot readily be used to detect small differences in enantioselectivity.

9.4 Troubleshooting

9.4.1 Comments on the Kazlauskas Test

The high-throughput Kazlauskas test for enantioselective hydrolases is inexpensive and practical. As noted by the author, true *E* values in the kinetic resolution of chiral esters are not provided, because the (*S*) and (*R*) substrates are tested separately [15a]. However, the relative initial rates provide an estimate of enantioselectivity, and the hits can then be studied conventionally using the racemate in conjunction with standard analytical tools such as chiral GC or HPLC. Sometimes serious discrepancies arise [15b]. A related colorimetric assay makes use of a perhaps more practical and sensitive indicator (bromothymol blue) [15c]. Finally, it must be kept in mind that *ee* assays showing a precision between ± 10 and 20% in the *ee* value, as in these screens and in some others, are well suited to identifying hits in the early phases of a directed evolution project. However, higher precision (better than $\pm 10\%$) makes screening in the later stages much easier [11a].

9.4.2 Potential Problems when Performing Kinetic Resolution

When taking samples for the determination of enantiomeric purity in a kinetic resolution, it is mandatory to choose the proper time window, e. g., when conversion is between 20% and 80%. If the process of directed evolution leads to enzyme variants differing widely in activity, mass screening needs to be time-resolved, i.e., several samples of a given reaction should be taken as a function of time.

9.5 Conclusions

A number of high-throughput *ee* assays have been described in the literature, but none is universal. When choosing a given method, several factors need to be considered, including cost, degree of throughput, and precision. In the early phases of projects concerning the directed evolution of enantioselective enzymes, low precision suffices, but when evolving more selective mutants at later stages ($ee > 90\%$) precision better than $\pm 10\%$ is not only helpful, but may be necessary. Currently it appears that assays based on UV/visible spectroscopy, mass spectrometry, and NMR and IR spectroscopy are most efficient and practical. The development of selection systems for enantioselectivity remains a challenge for the future, phage display being one possible strategy.

References

1. a) E. N. Jacobsen, A. Pfaltz, H. Yamamoto, *Comprehensive Asymmetric Catalysis*, Vol. I–III, Springer, Berlin, **1999**. b) R. Noyori, *Asymmetric Catalysis in Organic Synthesis*, Wiley, New York, **1994**. c) K. D. Shimizu, M. L. Snapper, A. H. Hoveyda, *Chem. Eur. J.* **1998**, *4*, 1885–1889.
2. S. C. Stinson, *Chem. Eng. News* **2001**, *79(40)*, 79–97.
3. B. Cornils, W. A. Herrmann, *Applied Homogeneous Catalysis with Organometallic Compounds*, Vol. 1–2, VCH-Wiley, Weinheim, **1996**.
4. a) H. G. Davies, R. H. Green, D. R. Kelly, S. M. Roberts, *Biotransformations in Preparative Organic Chemistry: The Use of Isolated Enzymes and Whole Cell Systems in Synthesis*, Academic Press, London, **1989**. b) C. H. Wong, G. M. Whitesides, *Enzymes in Synthetic Organic Chemistry, Tetrahedron Organic Chemistry Series*, Vol. 12, Pergamon, Oxford, **1994**. c) K. Drauz, H. Waldmann, *Enzyme Catalysis in Organic Synthesis: A Comprehensive Handbook*, Vol. I–III, VCH-Wiley, Weinheim, **2002**. d) K. Faber, *Biotransformations in Organic Chemistry*, 3rd ed., Springer, Berlin, **1997**.
5. A. Liese, K. Seelbach, C. Wandrey, *Industrial Biotransformations*, VCH-Wiley, Weinheim, **2000**.
6. a) M. T. Reetz, A. Zonta, K. Schimossek, K. Liebeton, K.-E. Jaeger, *Angew. Chem.* **1997**, *109*, 2961–2963; *Angew. Chem. Int. Ed. Engl.* **1997**, *36*, 2830–2832. b) K. Liebeton, A. Zonta, K. Schimossek, M. Nardini, D. Lang, B. W. Dijkstra, M. T. Reetz, K.-E. Jaeger, *Chem. Biol.* **2000**, *7*, 709–718. c) M. T. Reetz, S. Wilensek, D. Zha, K.-E. Jaeger, *Angew. Chem.* **2001**, *113*, 3701–3703,*Angew. Chem. Int. Ed.* **2001**, *40*, 3589–3591. d) D. Zha, S. Wilensek, M. Hermes, K.-E. Jaeger, M. T. Reetz, *Chem. Commun. (Cambridge)* **2001**, 2664–2665. e) M. T. Reetz, *Pure Appl. Chem.* **2000**, *72*, 1615–1622. f) M. T. Reetz, K.-E. Jaeger, *Chem. Eur. J.* **2000**, *6*, 407–412. g) M. T. Reetz, *Tetrahedron* **2002**, *58*, 6595–6602.
7. a) U. T Bornscheuer, J. Altenbuchner, H. H. Meyer, *Biotechnol. Bioeng.* **1998**, *58*, 554–559. b) O. May, P. T. Nguyen, F. H. Arnold, *Nat. Biotechnol.* **2000**, *18*, 317–320. c) S. Fong, T. D. Machajewski, C. C. Mak, C.-H. Wong, *Chem. Biol.* **2000**, *7*, 873–883.

8. For example: a) S. F. Brady, C. J. Chao, J. Handelsman, J. Clardy, *Org. Lett.* **2001**, *3*, 1981–1984. b) P. Hugenholtz, B. M. Goebel, N. R. Pace, *J. Bacteriol.* **1998**, *180*, 4765–4774. c) S. B. Bintrim, T. J. Donohue, J. Handelsman, G. P. Roberts, R. M. Goodman, *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 277–282.
9. G. DeSantis, Z. Zhu, W. A. Greenberg, K. Wong, J. Chaplin, S. R. Hanson, B. Farwell, L. W. Nicholson, C. L. Rand, D. P. Weiner, D. E. Robertson, M. J. Burk, *J. Am. Chem. Soc.* **2002**, *124*, 9024–9025.
10. a) F. H. Arnold, *Nature* **2001**, *409*, 253–257. b) K. A. Powell, S. W. Ramer, S. B. del Cardayré, W. P. C. Stemmer, M. B. Tobin, P. F. Longchamp, G. W. Huisman, *Angew. Chem.* **2001**, *113*, 4068–4080; *Angew. Chem. Int. Ed.* **2001**, *40*, 3948–3959. c) J. D. Sutherland, *Curr. Opin. Chem. Biol.* **2000**, *4*, 263–269. d) S. Brakmann, K. Johnsson, *Directed Molecular Evolution of Proteins*, Wiley-VCH, Weinheim, **2002.** e) F. H. Arnold, G. Georgiou, *Directed Enzyme Evolution (Screening and Selection Methods),* Humana Press, Totowa, New Jersey, **2003**.
11. a) M. T. Reetz, *Angew. Chem.* **2001**, *113*, 292–320;*Angew. Chem. Int. Ed.* **2001**, *40*, 284–310. b) M. T. Reetz, *Angew. Chem.* **2002**, *114*, 1391–1394. c) D. Wahler, J.-L. Reymond, *Curr. Opin. Biotechnol.* **2001**, *12*, 535–544. d) S. Dahmen, S. Bräse, *Synthesis* **2001**, 1431–1449.
12. C.-S. Chen, Y. Fujimoto, G. Girdaukas, C. J. Sih, *J. Am. Chem. Soc.* **1982**, *104*, 7294–7299.
13. M. T. Reetz, C. J. Rüggeberg, *Chem. Commun. (Cambridge)* **2002**, 1428–1429.
14. L. E. Janes, R. J. Kazlauskas, *J. Org. Chem.* **1997**, *62*, 4560–4561.
15. a) L. E. Janes, A. C. Löwendahl, R. J. Kazlauskas, *Chem. Eur. J.* **1998**, *4*, 2324–2331. b) A. M. F. Liu, N. A. Somers, R. J. Kazlauskas, T. S. Brush, F. Zocher, M. M. Enzelberger, U. T. Bornscheuer, G. P. Horsman, A. Mezzetti, C. Schmidt-Dannert, R. D. Schmid, *Tetrahedron: Asymmetry* **2001**, *12*, 545–556. c) F. Morís-Varas, A. Shah, J. Aikens, N. P. Nadkarni, J. D. Rozzell, D. C. Demirjian, *Bioorg. Med. Chem.* **1999**, *7*, 2183–2188.
16. M. Baumann, R. Stürmer, U. T. Bornscheuer, *Angew. Chem.* **2001**, *113*, 4329–4333; *Angew. Chem. Int. Ed.* **2001**, *40*, 4201–4204.
17. P. Abato, C. T. Seto, *J. Am. Chem. Soc.* **2001**, *123*, 9206–9207.
18. F. Taran, C. Gauchet, B. Mohar, S. Meunier, A. Valleix, P. Y. Renard, C. Créminon, J. Grassi, A. Wagner, C. Mioskowski, *Angew. Chem.* **2002**, *114*, 132–135; *Angew. Chem. Int. Ed.* **2002**, *41*, 124–127.
19. J. Guo, J. Wu, G. Siuzdak, M. G. Finn, *Angew. Chem.* **1999**, *111*, 1868–1871; *Angew. Chem. Int. Ed.* **1999**, *38*, 1755–1758.
20. a) M. T. Reetz, M. H. Becker, H.-W. Klein, D. Stöckigt, *Angew. Chem.* **1999**, *111*, 1872–1875; *Angew. Chem. Int. Ed.* **1999**, *38*, 1758–1761. b) W. Schrader, A. Eipper, D. J. Pugh, M. T. Reetz, *Can. J. Chem.* **2002**, *80*, 626–632. c) M. T. Reetz, M. H. Becker, D. Stöckigt, H.-W. Klein, patent applications DE-A 19913858.3, PCT/EP 00/02121, and WO 00/58504.
21. a) M. T. Reetz, A. Eipper, P. Tielmann, R. Mynott, *Adv. Synth. Catal.* **2002**, *344*, 1008–1016. b) M. T. Reetz, P. Tielmann, A. Eipper, R. Mynott, patent application DE-A 10209177.3.
22. P. Tielmann, M. Boese, M. Luft, M. T. Reetz, *Chem. Eur. J.* **2003**, *9*, 3882–3887.
23. a) K. Ding, A. Ishii, K. Mikami, *Angew. Chem.* **1999**, *111*, 519–523; *Angew. Chem. Int. Ed.* **1999**, *38*, 497–501. b) R. Angelaud, Y. Matsumoto, T. Korenaga, K. Kudo, M. Senda, K. Mikami, *Chirality* **2000**, *12*, 544–547.
24. M. T. Reetz, K. M. Kühling, H. Hinrichs, A. Deege, *Chirality* **2000**, *12*, 479–482.
25. M. T. Reetz, K. M. Kühling, A. Deege, H. Hinrichs, D. Belder, *Angew. Chem.* **2000**, *112*, 4049–4052; *Angew. Chem. Int. Ed.* **2000**, *39*, 3891–3893.

26. a) G. Klein, J.-L. Reymond, *Helv. Chim. Acta* **1999**, *82*, 400–406. b) E. Leroy, N. Bensel, J.-L. Reymond, *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2105–2108. c) D. Lagarde, H.-K. Nguyen, G. Ravot, D. Wahler, J.-L. Reymond, G. Hills, T. Veit, F. Lefevre, *Org. Process Res. Dev.* **2002**, *6*, 441–445. d) F. Badalassi, D. Wahler, G. Klein, P. Crotti, J.-L. Reymond, *Angew. Chem.* **2000**, *112*, 4233–4236; *Angew. Chem. Int. Ed.* **2000**, *39*, 4067–4077. e) J.-L. Reymond, D. Wahler, *ChemBioChem* **2002**, *3*, 701–708.
27. E. R. Jarvo, C. A. Evans, G. T. Copeland, S. J. Miller, *J. Org. Chem.* **2001**, *66*, 5522–5527.
28. G. A. Korbel, G. Lalic, M. D. Shair, *J. Am. Chem. Soc.* **2001**, *123*, 361–362.
29. M. T. Reetz, K. M. Kühling, S. Wilensek, H. Husmann, U. W. Häusig, M. Hermes, *Catal. Today* **2001**, *67*, 389–396.
30. a) M. T. Reetz, M. H. Becker, K. M. Kühling, A. Holzwarth, *Angew. Chem.* **1998**, *110*, 2792–2795; *Angew. Chem. Int. Ed.* **1998**, *37*, 2647–2650. b) M. T. Reetz, M. Hermes, M. H. Becker, *Appl. Microbiol. Biotechnol.* **2001**, *55*, 531–536.
31. N. Millot, P. Borman, M. S. Anson, I. B. Campbell, S. J. F. Macdonald, M. Mahmoudian, *Org. Process Res. Dev.* **2002**, *6*, 463–470.
32. A. Horeau, A. Nouaille, *Tetrahedron Lett.* **1990**, *31*, 2707–2710.
33. S. A. Funke, A. Eipper, M. T. Reetz, N. Otte, W. Thiel, G. van Pouderoyen, B. W. Dijkstra, K.-E. Jaeger, T. Eggert, *Biocatal. Biotransform.* **2003**, *21*, 67–73.
34. D. Zha, A. Eipper, M. T. Reetz, *ChemBioChem* **2003**, *4*, 34–39.
35. a) F. Cedrone, S. Niel, S. Roca, T. Bhatnagar, N. Ait-Abdelkader, C. Torre, H. Krumm, A. Maichele, M. T. Reetz, J. C. Baratti, *Biocatal. Biotransform.* **2003**, *21*, 357–364. b) M. T. Reetz, C. Torre, A. Eipper, R. Lohmer, M. Hermes, B. Brunner, A. Maichele, M. Bocola, M. Arand, A. Cronin, Y. Genzel, A. Archelas, R. Furstoss, *Org. Lett.* **2004**, *6*, 177–180.
36. G. DeSantis, K. Wong, B. Farwell, K. Chatman, Z. Zhu, G. Tomlinson, H. Huang, X. Tan, L. Bibbs, P. Chen, K. Kretz, and M. J. Burk, *J. Am. Chem. Soc.* **2003**, *125*, 11476–11477.
37. T. Eggert, G. Pencreac'h, I. Douchet, R. Verger, K.-E. Jaeger, *Eur. J. Biochem.* **2000**, *267*, 6459–6469.
38. a) M. J. Shapiro, J. S. Gounarides, *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *35*, 153–200. b) H. Schröder, P. Neidig, G. Rossé, *Angew. Chem.* **2000**, *112*, 3974–3977; *Angew. Chem. Int. Ed.* **2000**, *39*, 3816–3819. c) C. L. Gavaghan, J. K. Nicholson, S.C. Connor, I.D. Wilson, B. Wright, E. Holmes, *Anal. Biochem.* **2001**, *291*, 245–252. d) E. MacNamara, T. Hou, G. Fisher, S. Williams, D. Raftery, *Anal. Chim. Acta* **1999**, *397*, 9–16.
39. N. Krebsfänger, F. Zocher, J. Altenbuchner, U. T. Bornscheuer, *Enzyme Microb. Technol.* **1998**, *21*, 641–646.
40. A. Holzwarth, H.-W. Schmidt, W. F. Maier, *Angew. Chem.* **1998**, *110*, 2788–2792; *Angew. Chem. Int. Ed*. **1998**, *37*, 2644–2647.

10 Computer-assisted Design of Doped Libraries

Dirk Tomandl and Andreas Schwienhorst

10.1 Introduction

In general, the aim of directed evolution is to select molecules with desired molecular properties from a huge, diverse molecular repertoire. Ideally, this repertoire should comprise as many different molecular species as possible, to increase the probability of finding a molecular solution to the set optimization problem. However, combinatorial libraries of biopolymers that comprise all possible variants of a given length easily exceed the number of molecules that can be dealt with in a laboratory experiment. For a library of all possible 20-mer peptides, there are 20^{20} different variants with a total weight of two tons of material. Typical genetic selection systems can cope with much less than that. For example, phage display systems [1, 2] today can deal with up to 10^{11} molecules, which corresponds to a complete library of all possible octapeptides. Furthermore, for random libraries one has to consider that molecules with desired properties are usually highly diluted in a huge background of nonfunctional molecules. Single, functional molecules, therefore, can be easily missed by the selection procedure applied.

A feasible way to increase the fraction of functional molecules is to employ some *a priori* information, such as physicochemical parameters or phylogenetic information, to restrict the set of all possible building blocks at a certain sequential position to a subset of 'promising' monomers, i. e., 'doping'. In this way, the number of 'randomized' positions in a given sequence can be increased without exceeding the limits of experimentally feasible library size. Concerning protein libraries, doping would reveal molecules with only a certain subset of amino acids at a given 'randomized' position, i. e., a subset of codons at the corresponding position in the coding DNA. Within the scope of such doping strategies, it is important to avoid stop codons [3, 4] and to apply a codon usage that supports good expression in the expression organism, e. g., *Escherichia coli* [5–10].

To generate a protein library, partially randomized coding DNA has to be chemically synthesized and cloned into the context of a protein-encoding gene as part of a suitable expression vector, e. g., by using cassette mutagenesis [11]. Ideally, coding DNA is synthesized codon-wise, i. e., from trinucleotide building blocks [12]. However, since neither trimer building blocks nor corresponding synthesizers are commercially available, only a minority of researchers have access to this demanding technology. As a way out, each codon mixture can be realized on

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst
Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-30799-0

the basis of mononucleotide mixtures for each position of a partially randomized codon.

Here we provide a computer program that 'reverse translates' a desired target set of amino acids (for a given 'doped' position in a protein) into three mixtures of nucleotide synthons for each nucleotide position in the corresponding coding DNA sequence. The method is based on a hybrid algorithm (GALO) comprising a genetic algorithm and a local optimization method [13]. The fitness of the solutions is assessed by calculating the mean of the squared differences (which corresponds to the mean of the squared errors, MSE [13]) of single amino acid fractions from the desired values. Alternatively, the sum of absolute errors (SAE) is calculated.

For many problems, rather different solutions of similar fitness exist. To obtain a quick overview of the structure of the underlying high-dimensional fitness landscape, we used a self-organizing map (SOM) that was developed originally by Kohonen [14] in this study. Figure 10.1 shows a nonlinear mapping of a seven-

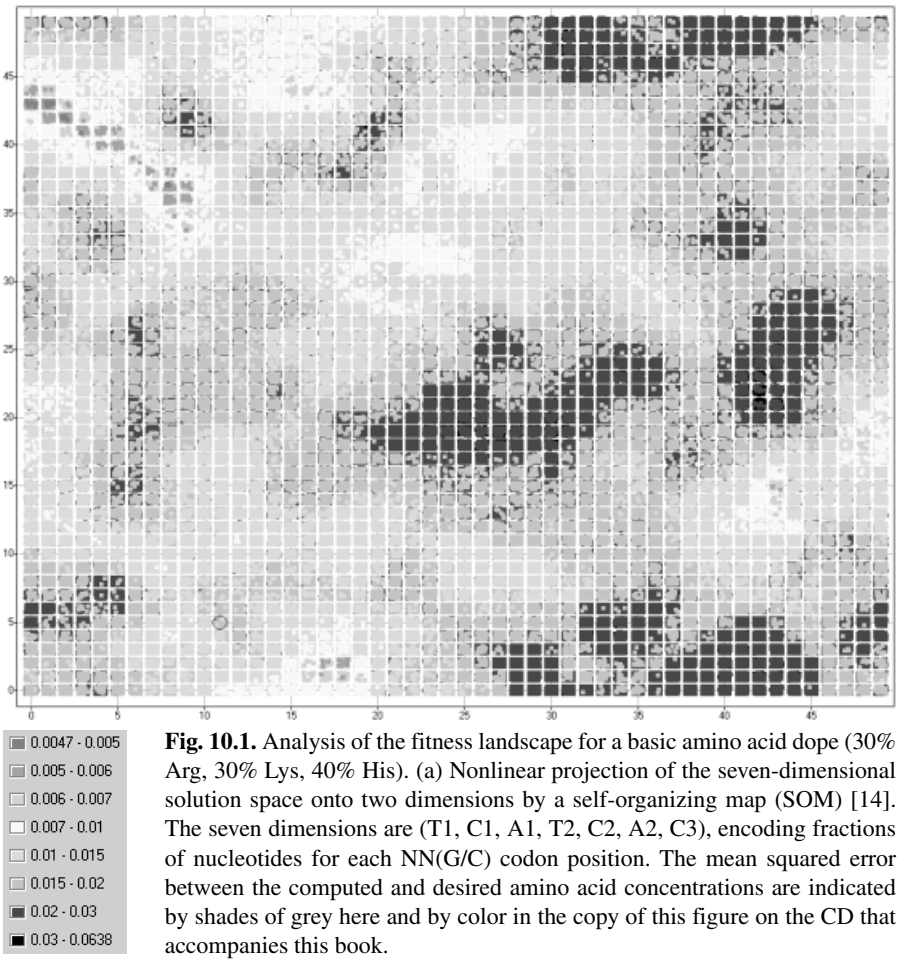


Fig. 10.1. Analysis of the fitness landscape for a basic amino acid dope (30% Arg, 30% Lys, 40% His). (a) Nonlinear projection of the seven-dimensional solution space onto two dimensions by a self-organizing map (SOM) [14]. The seven dimensions are (T1, C1, A1, T2, C2, A2, C3), encoding fractions of nucleotides for each NN(G/C) codon position. The mean squared error between the computed and desired amino acid concentrations are indicated by shades of grey here and by color in the copy of this figure on the CD that accompanies this book.

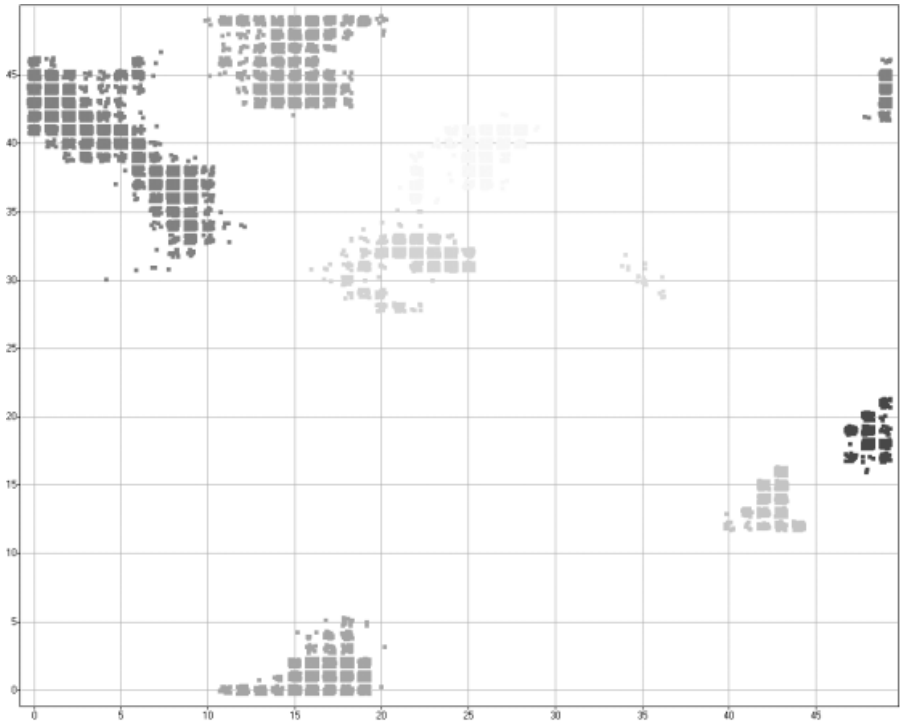
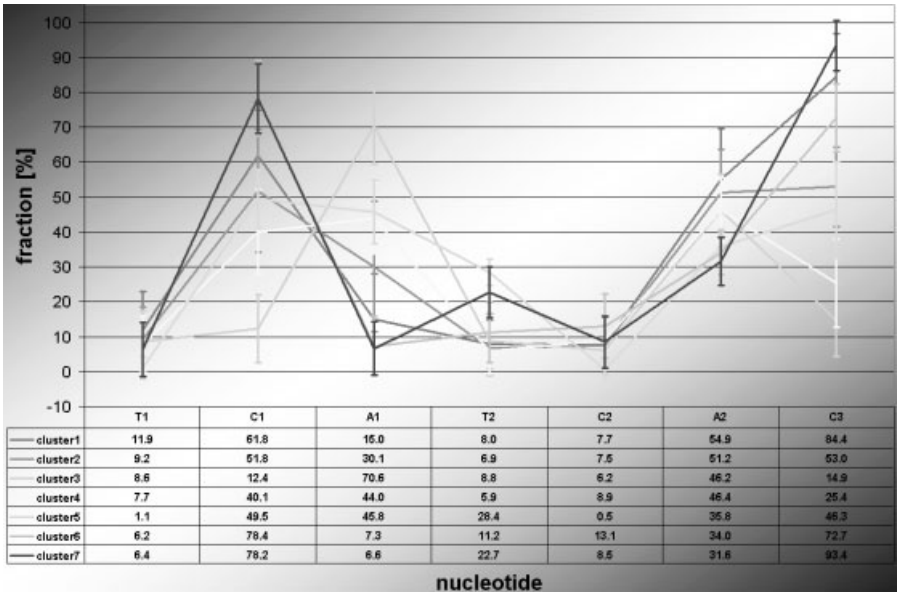


Fig. 10.1 (Continued). (b) Distribution of local optima in solution space for a basic amino acid dope (30% Arg, 30% Lys, 40% His). Due to the structure of the genetic code, a perfect solution of fractions of nucleotides for the given example does not exist. Instead, several islands of different suboptimal solutions are found. Seven local optima, marked by different colors (see the color version on the CD that accompanies this book), can be identified. Note that the SOM uses toroidal boundaries, i. e., the left-most and right-most points lie close together, as do the top and bottom points.

dimensional solution space to two dimensions for a basic amino acid dope (30% Arg, 30% Lys, 40% His). Here, seven different optima of similar fitness exist.

At least for the examples the authors studied, the fitness landscapes seem to be in general well suited for optimization algorithms. They consist of a rather low number of smooth local optima. Within single optima usually many similar solutions build a network of ridges, facilitating algorithmic optimization and lending the optima a certain robustness against small experimental errors in the absolute synthon concentrations [13].

In general, libraries are created by synthesizing a single doped oligonucleotide (‘one-pot synthesis’). Sometimes, better solutions can be obtained by synthesizing two or three different doped oligonucleotides (‘two-pot’ or ‘three-pot’ synthesis) and preparing a mixture of these oligonucleotides.



	T1	C1	A1	T2	C2	A2	C3
cluster1	11.9	61.8	15.0	8.0	7.7	54.9	84.4
cluster2	9.2	51.8	30.1	6.9	7.5	51.2	53.0
cluster3	8.6	12.4	70.6	8.8	6.2	46.2	14.9
cluster4	7.7	40.1	44.0	5.9	8.9	46.4	25.4
cluster5	1.1	49.5	45.8	28.4	0.5	35.8	46.3
cluster6	6.2	78.4	7.3	11.2	13.1	34.0	72.7
cluster7	6.4	78.2	6.6	22.7	8.5	31.6	93.4

Fig. 10.1 (Continued). (c) Distribution of solutions within each local optimum for a basic amino acid dope (30% Arg, 30% Lys, 40% His). Each cluster consists of many different solutions, which are overlaid in the profile plots, revealing some trends. Each cluster emphasizes a different typical profile of solutions – and clusters 2 and 5 are especially different from the other clusters. Comparing the mean values of each nucleotide between the clusters shows large differences, especially for the nucleotide fractions A1 and C3. The standard deviation for each nucleotide is ∼10%.

The computer program presented here can generate very different relative frequencies of occurrence of all amino acids (and stop codons) within a target set. One-pot, two-pot, and three-pot syntheses can be simulated. Optionally, correction factors can be included to compensate for the (possible) differences in chemical coupling efficiencies of the nucleotide synthons. Finally, the codon usage of three major expression microorganisms can be considered, treating rare codons as pseudo stop codons.

10.2 Materials

The computer tool consists of two .exe files that implement the GALO algorithm and run under 32-bit MS DOS, a .pl file that implements the graphical user interface (GUI), and three codon-usage files. The GUI requires an installed Perl5 distribution, including the Perl/Tk modules (see, e. g., www.activestate.com). The GUI provides

the necessary input and output fields and serves as a convenient 'shell' around the command line .exe files, DOPING_G.EXE and DOPING_A.EXE. The .exe files, however, can be run as standalone tools as well. Before using them, copy all program and codon-usage files onto the hard disk of your computer. Then simply double-click the file DOPING.PL to start the GUI.

Minimal system requirements are those required for Windows 95 and Perl/Tk. A graphical resolution of at least 1024 × 768 pixels with 256 colors is recommended, because the GUI window is rather large (about 900 × 700 pixels).

If other codon-usage tables will be used, just download the desired codon usage files from www.kazusa.or.jp/codon and add a first line containing the threshold below which codons should be regarded as pseudo stop codons.

10.3 Protocol

The starting point of almost all experiments in directed protein evolution is to generate a molecular repertoire on the level of DNA. In general, DNA libraries are synthesized from mononucleotide phosphoramidite building blocks. Partially randomized amino acid positions in a protein sequence are then created on the basis of mononucleotide mixtures for each position of a partially randomized codon in the coding DNA. Our computer program reverse translates a desired target set of amino acids into three mixtures of nucleotide synthons for each nucleotide position in the corresponding codon (GUI is shown in Figure 10.2). Although the procedure is largely self-explanatory, a brief overview is given below.

1. First, enter the desired fractions of amino acids (as percentages) in the white input boxes on the left side of the GUI.
2. Next, select the number of synthesis pots (default is one-pot synthesis). You can also constrain the algorithm to only use G and C nucleotides in the third codon position. These controls are in the lower right of the GUI.
3. Nucleotide fractions appear in the boxes on the right side of the GUI. They can be entered manually or calculated by clicking the *Compute by GALO* button at the bottom of the GUI. In the current version of the GALO method [13], nucleotide fractions are calculated that are multiples of 1% of the total number of synthons in the mixture. Nucleotide fractions can be edited at any time. Click the *Compute Manually* button to recalculate the fractions of amino acids.
   The corresponding codon distributions (as percentages) are shown in the light blue boxes. For multiple-pot synthesis, the fraction of each individual doped oligonucleotide in the final mixture is also given.
4. (optional) If a certain codon usage has to be considered, select the corresponding table from the pulldown menu in the lower right of the GUI. All codons now regarded as stop codons are colored light green. To recalculate nucleotide fractions click the *Compute by GALO* or *Compute Manually* button again.

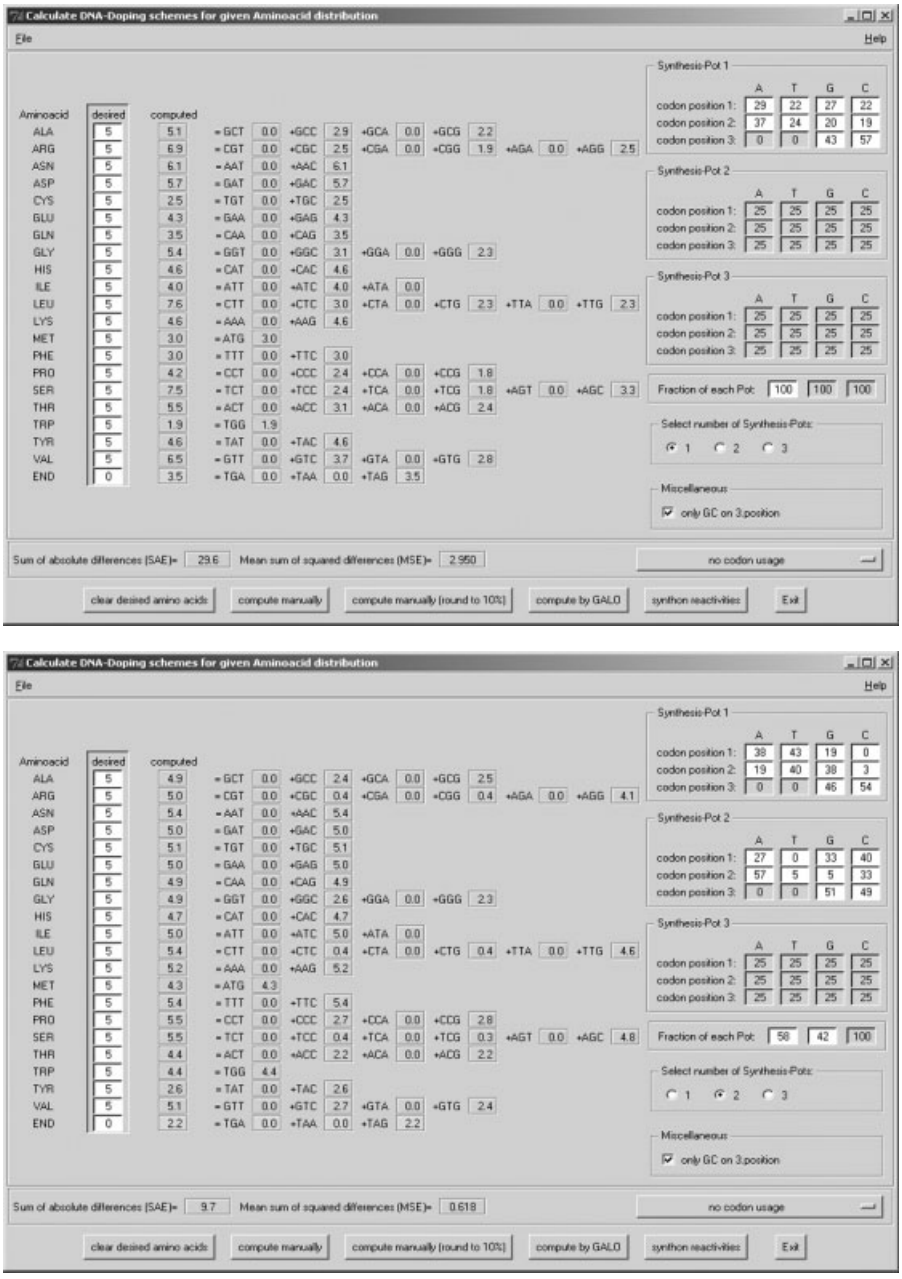

Calculate DNA-Doping schemes for given Aminoacid distribution
File
Help
Aminoacid
desired
computed
ALA 5 5.1 = GCT 0.0 +GCC 2.9 +GCA 0.0 +GCG 2.2
ARG 5 6.9 = CGT 0.0 +CGC 2.5 +CGA 0.0 +CGG 1.9 +AGA 0.0 +AGG 2.5
ASN 5 6.1 = AAT 0.0 +AAC 6.1
ASP 5 5.7 = GAT 0.0 +GAC 5.7
CYS 5 2.5 = TGT 0.0 +TGC 2.5
GLU 5 4.3 = GAA 0.0 +GAG 4.3
GLN 5 3.5 = CAA 0.0 +CAG 3.5
GLY 5 5.4 = GGT 0.0 +GGC 3.1 +GGA 0.0 +GGG 2.3
HIS 5 4.6 = CAT 0.0 +CAC 4.6
ILE 5 4.0 = ATT 0.0 +ATC 4.0 +ATA 0.0
LEU 5 7.6 = CTT 0.0 +CTC 3.0 +CTA 0.0 +CTG 2.3 +TTA 0.0 +TTG 2.3
LYS 5 4.6 = AAA 0.0 +AAG 4.6
MET 5 3.0 = ATG 3.0
PHE 5 3.0 = TTT 0.0 +TTC 3.0
PRO 5 4.2 = CCT 0.0 +CCC 2.4 +CCA 0.0 +CCG 1.8
SER 5 7.5 = TCT 0.0 +TCC 2.4 +TCA 0.0 +TCG 1.8 +AGT 0.0 +AGC 3.3
THR 5 5.5 = ACT 0.0 +ACC 3.1 +ACA 0.0 +ACG 2.4
TRP 5 1.9 = TGG 1.9
TYR 5 4.6 = TAT 0.0 +TAC 4.6
VAL 5 6.5 = GTT 0.0 +GTC 3.7 +GTA 0.0 +GTG 2.8
END 0 3.5 = TGA 0.0 +TAA 0.0 +TAG 3.5
Synthesis-Pot 1
A T G C
codon position 1: 29 22 27 22
codon position 2: 37 24 20 19
codon position 3: 0 0 43 57
Synthesis-Pot 2
A T G C
codon position 1: 25 25 25 25
codon position 2: 25 25 25 25
codon position 3: 25 25 25 25
Synthesis-Pot 3
A T G C
codon position 1: 25 25 25 25
codon position 2: 25 25 25 25
codon position 3: 25 25 25 25
Fraction of each Pot: 100 100 100
Select number of Synthesis-Pots:
1 2 3
Miscellaneous
only GC on 3.position
Sum of absolute differences (SAE)= 29.6
Mean sum of squared differences (MSE)= 2.950
no codon usage
clear desired amino acids
compute manually
compute manually (round to 10%)
compute by GALO
synthon reactivities
Exit
Calculate DNA-Doping schemes for given Aminoacid distribution
File
Help
Aminoacid
desired
computed
ALA 5 4.9 = GCT 0.0 +GCC 2.4 +GCA 0.0 +GCG 2.5
ARG 5 5.0 = CGT 0.0 +CGC 0.4 +CGA 0.0 +CGG 0.4 +AGA 0.0 +AGG 4.1
ASN 5 5.4 = AAT 0.0 +AAC 5.4
ASP 5 5.0 = GAT 0.0 +GAC 5.0
CYS 5 5.1 = TGT 0.0 +TGC 5.1
GLU 5 5.0 = GAA 0.0 +GAG 5.0
GLN 5 4.9 = CAA 0.0 +CAG 4.9
GLY 5 4.9 = GGT 0.0 +GGC 2.6 +GGA 0.0 +GGG 2.3
HIS 5 4.7 = CAT 0.0 +CAC 4.7
ILE 5 5.0 = ATT 0.0 +ATC 5.0 +ATA 0.0
LEU 5 5.4 = CTT 0.0 +CTC 0.4 +CTA 0.0 +CTG 0.4 +TTA 0.0 +TTG 4.6
LYS 5 5.2 = AAA 0.0 +AAG 5.2
MET 5 4.3 = ATG 4.3
PHE 5 5.4 = TTT 0.0 +TTC 5.4
PRO 5 5.5 = CCT 0.0 +CCC 2.7 +CCA 0.0 +CCG 2.8
SER 5 5.5 = TCT 0.0 +TCC 0.4 +TCA 0.0 +TCG 0.3 +AGT 0.0 +AGC 4.8
THR 5 4.4 = ACT 0.0 +ACC 2.2 +ACA 0.0 +ACG 2.2
TRP 5 4.4 = TGG 4.4
TYR 5 2.6 = TAT 0.0 +TAC 2.6
VAL 5 5.1 = GTT 0.0 +GTC 2.7 +GTA 0.0 +GTG 2.4
END 0 2.2 = TGA 0.0 +TAA 0.0 +TAG 2.2
Synthesis-Pot 1
A T G C
codon position 1: 38 43 19 0
codon position 2: 19 40 38 3
codon position 3: 0 0 46 54
Synthesis-Pot 2
A T G C
codon position 1: 27 0 33 40
codon position 2: 57 5 5 33
codon position 3: 0 0 51 49
Synthesis-Pot 3
A T G C
codon position 1: 25 25 25 25
codon position 2: 25 25 25 25
codon position 3: 25 25 25 25
Fraction of each Pot: 58 42 100
Select number of Synthesis-Pots:
1 2 3
Miscellaneous
only GC on 3.position
Sum of absolute differences (SAE)= 9.7
Mean sum of squared differences (MSE)= 0.618
no codon usage
clear desired amino acids
compute manually
compute manually (round to 10%)
compute by GALO
synthon reactivities
Exit

Fig. 10.2. Screenshots of the GUI after obtaining results for the problem of equimolar mixtures of amino acids, i. e., 5% of each amino acid and 0% of stop codons. (a) Standard one-pot synthesis; (b) two-pot synthesis; (c) three-pot synthesis; (d) one-pot synthesis using *E. coli* codon usage. Input fields are white and inactive fields are light gray. Computed fields are light blue and stop codons are light green (in the copy of this figure on the CD that accompanies this book).

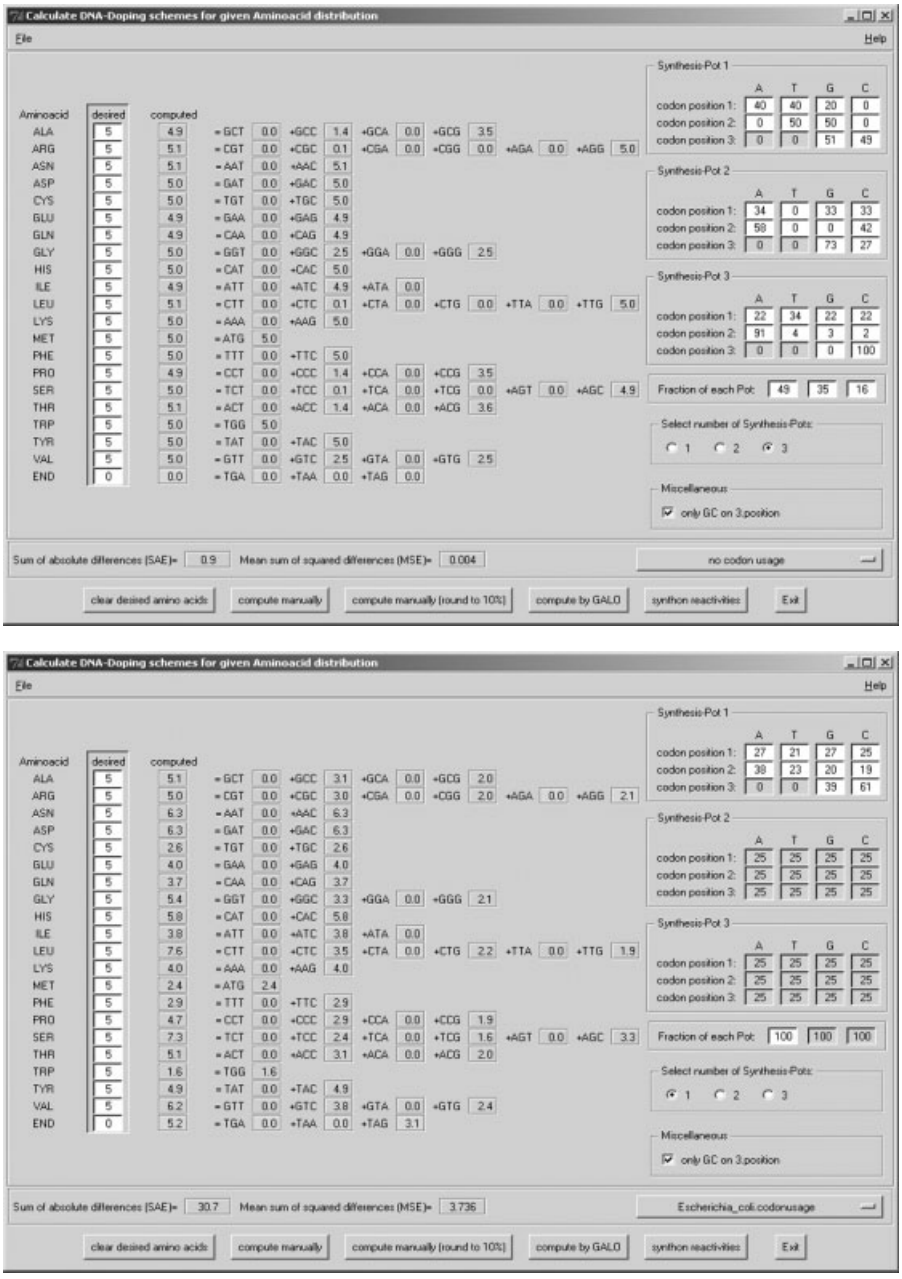

Calculate DNA-Doping schemes for given Aminoacid distribution
File
Help
Aminoacid
desired
computed
ALA 5 4.9 = GCT 0.0 +GCC 1.4 +GCA 0.0 +GCG 3.5
ARG 5 5.1 = CGT 0.0 +CGC 0.1 +CGA 0.0 +CGG 0.0 +AGA 0.0 +AGG 5.0
ASN 5 5.1 = AAT 0.0 +AAC 5.1
ASP 5 5.0 = GAT 0.0 +GAC 5.0
CYS 5 5.0 = TGT 0.0 +TGC 5.0
GLU 5 4.9 = GAA 0.0 +GAG 4.9
GLN 5 4.9 = CAA 0.0 +CAG 4.9
GLY 5 5.0 = GGT 0.0 +GGC 2.5 +GGA 0.0 +GGG 2.5
HIS 5 5.0 = CAT 0.0 +CAC 5.0
ILE 5 4.9 = ATT 0.0 +ATC 4.9 +ATA 0.0
LEU 5 5.1 = CTT 0.0 +CTC 0.1 +CTA 0.0 +CTG 0.0 +TTA 0.0 +TTG 5.0
LYS 5 5.0 = AAA 0.0 +AAG 5.0
MET 5 5.0 = ATG 5.0
PHE 5 5.0 = TTT 0.0 +TTC 5.0
PRO 5 4.9 = CCT 0.0 +CCC 1.4 +CCA 0.0 +CCG 3.5
SER 5 5.0 = TCT 0.0 +TCC 0.1 +TCA 0.0 +TCG 0.0 +AGT 0.0 +AGC 4.9
THR 5 5.1 = ACT 0.0 +ACC 1.4 +ACA 0.0 +ACG 3.6
TRP 5 5.0 = TGG 5.0
TYR 5 5.0 = TAT 0.0 +TAC 5.0
VAL 5 5.0 = GTT 0.0 +GTC 2.5 +GTA 0.0 +GTG 2.5
END 0 0.0 = TGA 0.0 +TAA 0.0 +TAG 0.0
Synthesis-Pot 1
A T G C
codon position 1: 40 40 20 0
codon position 2: 0 50 50 0
codon position 3: 0 0 51 49
Synthesis-Pot 2
A T G C
codon position 1: 34 0 33 33
codon position 2: 58 0 0 42
codon position 3: 0 0 73 27
Synthesis-Pot 3
A T G C
codon position 1: 22 34 22 22
codon position 2: 91 4 3 2
codon position 3: 0 0 0 100
Fraction of each Pot: 49 35 16
Select number of Synthesis-Pots:
1 2 3
Miscellaneous
only GC on 3.position
Sum of absolute differences (SAE)= 0.9
Mean sum of squared differences (MSE)= 0.004
no codon usage
clear desired amino acids
compute manually
compute manually (round to 10%)
compute by GALO
synthon reactivities
Exit
Calculate DNA-Doping schemes for given Aminoacid distribution
File
Help
Aminoacid
desired
computed
ALA 5 5.1 = GCT 0.0 +GCC 3.1 +GCA 0.0 +GCG 2.0
ARG 5 5.0 = CGT 0.0 +CGC 3.0 +CGA 0.0 +CGG 2.0 +AGA 0.0 +AGG 2.1
ASN 5 6.3 = AAT 0.0 +AAC 6.3
ASP 5 6.3 = GAT 0.0 +GAC 6.3
CYS 5 2.6 = TGT 0.0 +TGC 2.6
GLU 5 4.0 = GAA 0.0 +GAG 4.0
GLN 5 3.7 = CAA 0.0 +CAG 3.7
GLY 5 5.4 = GGT 0.0 +GGC 3.3 +GGA 0.0 +GGG 2.1
HIS 5 5.8 = CAT 0.0 +CAC 5.8
ILE 5 3.8 = ATT 0.0 +ATC 3.8 +ATA 0.0
LEU 5 7.6 = CTT 0.0 +CTC 3.5 +CTA 0.0 +CTG 2.2 +TTA 0.0 +TTG 1.9
LYS 5 4.0 = AAA 0.0 +AAG 4.0
MET 5 2.4 = ATG 2.4
PHE 5 2.9 = TTT 0.0 +TTC 2.9
PRO 5 4.7 = CCT 0.0 +CCC 2.9 +CCA 0.0 +CCG 1.9
SER 5 7.3 = TCT 0.0 +TCC 2.4 +TCA 0.0 +TCG 1.6 +AGT 0.0 +AGC 3.3
THR 5 5.1 = ACT 0.0 +ACC 3.1 +ACA 0.0 +ACG 2.0
TRP 5 1.6 = TGG 1.6
TYR 5 4.9 = TAT 0.0 +TAC 4.9
VAL 5 6.2 = GTT 0.0 +GTC 3.8 +GTA 0.0 +GTG 2.4
END 0 5.2 = TGA 0.0 +TAA 0.0 +TAG 3.1
Synthesis-Pot 1
A T G C
codon position 1: 27 21 27 25
codon position 2: 38 23 20 19
codon position 3: 0 0 39 61
Synthesis-Pot 2
A T G C
codon position 1: 25 25 25 25
codon position 2: 25 25 25 25
codon position 3: 25 25 25 25
Synthesis-Pot 3
A T G C
codon position 1: 25 25 25 25
codon position 2: 25 25 25 25
codon position 3: 25 25 25 25
Fraction of each Pot: 100 100 100
Select number of Synthesis-Pots:
1 2 3
Miscellaneous
only GC on 3.position
Sum of absolute differences (SAE)= 30.7
Mean sum of squared differences (MSE)= 3.736
Escherichia_coli.codonusage
clear desired amino acids
compute manually
compute manually (round to 10%)
compute by GALO
synthon reactivities
Exit

Fig. 10.2 (Continued).

5. (optional) If different reactivities of synthons have to be considered, click the *Synthon Reactivities* button. A second box opens in which you enter the relative reactivities manually. To recalculate the nucleotide fractions, click the *Apply Reaction Rates* button.

Figure 10.2 shows four screen shots of the GUIs for the problem of equimolar mixtures of amino acids, i. e., 5% of each amino acid and 0% of stop codons. In Figures 10.2a–c the results of a standard one-pot, two-pot, and three-pot synthesis are given, respectively. Figure 10.2d shows the result of a one-pot synthesis optimized for *E. coli* codon usage.

10.4 Troubleshooting

There is conflicting evidence about the reaction rates of the four standard phosphoramidites used in automated DNA synthesis (reviewed in [13]). Whereas some publications suggest that freshly prepared, nominally equimolar, premade mixtures of the amidites afford product distributions that are nearly equimolar (±1%–5%), others have reported that the A phosphoramidite is incorporated at a somewhat higher frequency than C, G, or T. In contradiction, the User Bulletin from Applied Biosystems states that the A phosphoramidite is the less reactive one. The same reference also mentions that the G phosphoramidite degrades faster than the others, making the age of the phosphoramidite solutions a critical point. Since the effects of different reaction rates have a tremendous influence on an experimentally generated dope, it seems essential to quantify them carefully for the synthesizer and the synthesis protocol of choice. If phosphoramidite reactivities indeed turn out to be significantly different, you should apply step 5 (above), i. e., calculate the necessary relative phosphoramidite concentrations to yield the desired nucleotide concentrations at each codon position.

10.5 Major Applications

The doping algorithm has already been useful in a number of examples. The dope for equimolar mixtures of all 20 amino acids optimized for the codon usage of *E. coli* was recently applied to the generation of hirudin gene libraries [15]. Here, the goal was to identify protease-resistant variants of this thrombin-specific inhibitor by using a combination of phage-display selection and high-throughput screening methods.

Phylogenetic data from 18 sequences of the thioredoxin family were used to calculate the doping scheme for active site positions [13]. The frequency of different amino acids at a given position determines a positional mixture with different

fractions of these amino acids, which is then reverse-translated into three nucleotide mixtures of the concomitant (doped) codon. Future applications may include not only phylogenetic but also physicochemical information from comparison of natural or artificially selected amino acid sequences, so as to increase the number of functional molecules in a random library.

Acknowledgments

This work was supported in part by grant BioFuture 0311852 from the Bundesministerium für Forschung und Technologie, Germany. We also thank A. Schober and F. Wirsching for discussions.

References

1. Smith, G. (**1985**) *Science 228*, 1315–1317.
2. Scott, J.K., and Smith, G.P. (**1990**) *Science 249*, 386–390.
3. Little, J.W. (**1990**) *Gene 88*, 113–115.
4. Siderovski, D.P., and Mak, T.W. (**1993**) *Comput. Biol. Med. 23*, 463–474.
5. Robinson, M., Lilley, R., Little, S., Emtage, J.S., Yarranton, G., Stephens, P., Millican, A., Eaton, M., and Humphreys, G. (**1984**) *Nucleic Acids Res. 12*, 6663–6671.
6. Spanjaard, R.A., and Van Duin, J. (**1988**) *Proc. Natl. Acad. Sci. USA 85*, 7967–7971.
7. Ernst, J.F., and Kawashima, E. (**1988**) *J. Biotechnol. 8*, 1–10.
8. Bonekamp, F., and Jensen, K.F. (**1988**) *Nucleic Acids Res. 16*, 3013–3024.
9. Brinkmann, U., Mattes, R.E., and Buckel, P. (**1989**) *Gene 85,* 109–114.
10. Schenk, P.M., Baumann, S., Mattes, R., and Steinbiss, H.H. (**1995**) *BioTechniques 19*, 196–200.
11. Borrego, B., Wienecke, A., and Schwienhorst, A. (**1995**) *Nucleic Acids Res. 23*, 1834–1835.
12. Virnekäs, B., Ge, L., Plückthun, A., Schneider, K.C., Wellnhofer, G., and Moroney, S. (**1994**) *Nucleic Acids Res. 22*, 5600–5607.
13. Tomandl, D., Schober, A., and Schwienhorst, A. (**1997**) *J. Comput. Aided Mol. Design 11*, 29–38.
14. Kohonen, T. (**1997**) *Self-organizing Maps*, Vol. 30, 2nd ed., Springer, Heidelberg.
15. Wirsching, F., Keller, M., Hildmann, C., Riester, D., and Schwienhorst, A. (**2003**) *Mol. Gen. Metab. 80*, 451–462.

11 Directed *in silico* Mutagenesis

Markus Wiederstein, Peter Lackner, Ferry Kienberger, Manfred J. Sippl

11.1 Introduction

Proteins evolve and adapt to specific biological roles by point mutations, deletions, insertions, and permutations of the DNA encoding their amino acid sequences. With this small repertoire of mechanisms, proteins are capable of evolving all the structures and functions we observe today in all of living nature. Protein engineering takes advantage of these mechanisms to create proteins with desired properties. The expectation is that at some point we will be able to design proteins with specific properties, and after expressing the proteins we will actually find these properties in our design. For this to come true we need an appropriate understanding of the folding and stability of proteins in their respective environments and, although we are not yet there, we have tools at hand already which help to estimate the stability of proteins and the effect of amino acid changes on a protein's properties. These tools are not perfect. One cannot apply them in a nonexpert environment nor to every problem we want to solve, but they do provide results and answers to certain types of problems that go beyond qualitative statements.

To be specific, let us imagine the following scenario, quite common in protein engineering. We start with a protein of known three–dimensional structure. Experimental results indicate that our protein is quite stable and well behaved. We want to use this protein as a starting point to 'evolve' specific features on the surface of the protein, that is, hydrophobic patches, cavities, binding sites, and the like, to achieve certain functions like binding and transport of a class of small molecules. A first step towards this goal is the identification of residue positions whose redesign might yield the desired results.

Even if we restrict our design to a small number of sites in the protein, the combinatorial possibilities quickly approach astronomical dimensions. If we consider mutations at 10 sites and a subset of 10 amino acids, we have 10^{10} possible variants. Although experimental approaches are under development that can actually search large subsets of protein sequence space, it is not at all a small feat to identify those variants that give rise to a stable structure and at the same time come close to the desired features. Therefore, computational approaches that, with some reliability, are able to pick those variants having a stable structure are desirable instruments in the protein engineer's toolbox.

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst
Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-30799-0

In what follows we give a brief introduction to the program package called ProSa (an acronym derived from protein structure analysis) and the knowledge-based potentials the program is built on, and we show how the program can be used to investigate the stability of proteins as a function of changes in amino acid sequences. The key term is 'stability' and we have to be clear about the meaning of this term before we proceed, since the word is used in a variety of contexts. The *thermodynamic stability* of a protein is usually defined in terms of the difference in Gibbs free energy:

$$\Delta G = G_{\text{unfolded}} - G_{\text{folded}} \tag{11.1}$$

and hence, the factors that determine the thermodynamic stability are entirely due to the folded (native) state and the unfolded (denatured) state of the protein. Most proteins have ΔG values in the range of 5 to 15 kcal mol^{-1}. The second major term is *kinetic stability*, which addresses the rate of folding. A kinetically stable protein unfolds more slowly than a kinetically unstable protein and the rate of folding or unfolding depends on the transition state and the associated energy barrier between the folded and unfolded states.

What we get from ProSa is a comparison of a given structure to a large number of alternative structures (that is, an ensemble) in terms of knowledge-based potentials and z-scores. These numbers are neither thermodynamic stabilities nor kinetic stabilities. The z-score tells us something about the confidence we have in the hypothesis that our structure resembles the native structure of our protein, given all the knowledge we have assembled from the database of known structures. A low z-score – scores are reported as negative numbers, so by 'low' we are referring to absolute values – means that there are many structures available to our sequence having the same knowledge-based energy (that is, the sum over all pair interactions and solvent interactions), and hence our confidence in the hypothesis that we have the native structure of our sequence is quite low. If the z-score is high (very negative), then it will be very difficult to find an alternative structure with the same (low) energy as our structure and hence we are rather confident that our structure is a good model for the native folding.

To get back to our protein engineering problem, let us start with a protein of known structure. We calculate the z-score and find this to be in the range of a typical native structure. We now start to change the amino acid sequence at one or more positions, but we do not change the structure. For each variant we get a new z-score due to the modified amino acid sequence. If the z-score becomes less significant then – from our experience gained on quite a few results obtained from our computations – we take this as an indication that the new sequence does not fit into the given structure as well as for the wild-type sequence. Since we are aware of the many approximations involved, the limited amount of data available, and the many complications that may or may not arise in our specific case, we are very cautious and assume this to be a valid statement in general and not in every particular instance.

On the other hand, if the z-score does not change, then we have some confidence in the hypothesis that our structure is also the native fold of the mutated sequence.

Finally, if the z-score increases, then this may indicate that the new sequence fits even better than the wild-type sequence. Intuitively these statements have a lot to do with the notion of stability. If the z-score decreases then the new sequence is less stable than our structure, and our reasoning resembles the thermodynamic definition of stability, but to investigate the relationship between z-scores and experimental ΔG values is a different matter, which we do not pursue here ([1] provides some data regarding this relationship).

ProSa builds on knowledge-based potentials. These are functions that are extracted from known protein structures and, as such, they describe the average distribution of atoms and atom–atom interactions in protein–solvent systems. Since they represent an average over a whole database of structures, details that might be relevant for a particular protein are averaged out. Therefore, the potentials provide us with the characteristics of an average protein as determined by X-ray crystallography. We have to remember that a crystallised protein is not necessarily in its native state (although that is what crystallographers often assert) and that the true native state is actually an ensemble of structures that more or less resemble the crystallised protein (this is how NMR spectroscopists often like to see an X-ray structure). Although this is an important point, it does not concern us too much here since our model of the protein–solvent system is incomplete. In ProSa we consider only C^{α} and C^{β} atoms, and the protein–solvent interactions are taken into account by a spherical shell used to estimate an amino acid's solvent exposure. Another important issue is the additivity or nonadditivity of pair interaction terms and solvent terms. Taking into account the approximations involved, we cannot expect to be very precise.

In what follows we provide a summary of ProSa's components (Figure 11.1) and present the protocols that we have developed for small- and large-scale protein engineering problems. The accompanying CD holds the latest version of ProSa containing an interface to access and use its functions.

11.2 Materials

11.2.1 PDB Files

ProSa reads all it needs to know about a protein from a file in PDB format. A detailed description of this standard format is given at the RCSB website [2]. Of all record types within a PDB file, only the ATOM records are used by ProSa. They contain both the amino acid sequence of a certain protein chain and the atomic coordinates (example PDB files come with the ProSa distribution). If a PDB file contains more than one chain, only one is read. The first chain is read by default, other chains can be specified by providing the respective chain identifier. All molecules other than polypeptides are ignored. For energy calculations, ProSa takes into account C^{α} and C^{β} atoms. Care has to be taken when dealing with residue

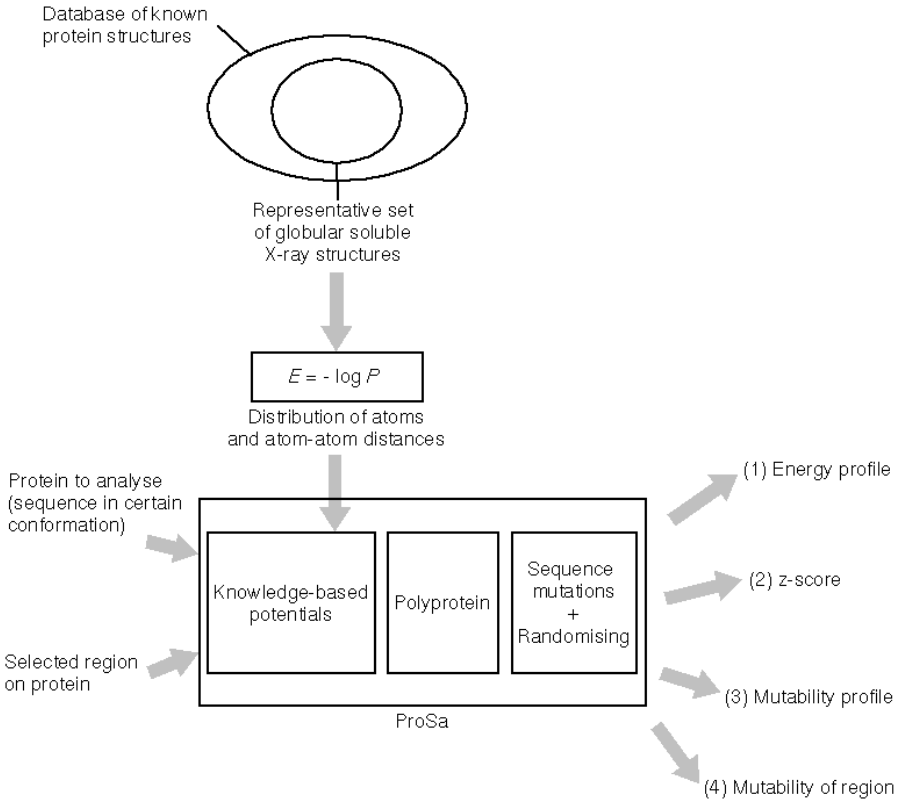


Fig. 11.1. ProSa components.

numbers. PDB format does not require that residue numbers start with 1, nor does it require consecutive residue numbering. ProSa ignores the numbering in the PDB file and assigns its own residue numbers, ranging from one to the length of the sequence. The command `print residue mapping` can be used to list PDB residue numbers and corresponding ProSa residue numbers.

11.2.2 Knowledge-based Potentials

In protein structure prediction, potentials are used to assign an energy-like quantity to a conformation of a protein molecule. If this quantity enables us to distinguish the native state of a protein, the potential is regarded as a reasonable model for a protein–solvent system. The rationale behind this relies on two assumptions: (a) a solved protein in its native state can be described by an ensemble of closely related conformations, and (b) in this state the system is in the global minimum of free energy. Virtually all techniques designed for structure prediction are based on these principles [3, 4].

In the knowledge-based approach presented here, the strategy is to derive potentials from known protein structures determined by X-ray diffraction. These structures are investigated with regard to the distributions of their atoms and atom–atom interactions. It is expected that these distributions contain information about typical native-like proteins. The next step is to relate these distributions to an energy-like quantity [1, 5–12].

ProSa uses two kinds of knowledge-based potentials, one to model pairwise interactions between protein atoms, and one for protein–solvent interactions.

11.2.2.1 Knowledge-based Potentials for Interactions between Protein Atoms

Let $\{a, b, c, d, k, r\}$ be a set of discrete variables that describe a particular conformation of a protein (see Table 11.1 and Figure 11.2). Then the knowledge-based potential for pair interactions is

$$E^{abcdk}(r) = -kT \ln f^{abcdk}(r) - kT \ln Z^{abcdk} \tag{11.2}$$

where the frequencies $f^{abcdk}(r)$ are obtained from the database as the relative frequencies of amino acid pairs (a, b) with atom types (c, d) and separation k along the sequence at spatial distance r. The partition function Z^{abcdk} remains undetermined but since it is constant, the term disappears when energy differences are calculated. To obtain a measure for the specific interactions of a particular amino acid pair, all information that is independent of the amino acid type has to be removed from $E^{abcdk}(r)$. This redundant information is captured by the average energy:

$$E^{cdk}(r) = -kT \ln f^{cdk}(r) - kT \ln Z^{cdk} \tag{11.3}$$

which is the knowledge-based potential for any pair of atoms of type c and d and separation k along the sequence at distance r. The net potential is then calculated by

$$\Delta E^{abcdk}(r) = E^{abcdk}(r) - E^{cdk}(r) = -kT \ln \left(\frac{f^{abcdk}(r)}{f^{cdk}(r)} \right) \tag{11.4}$$

Table 11.1. List of variables used for describing a protein conformation.

Variable	Description
a	amino acid
b	amino acid
c	atom type
d	atom type
r	spatial distance (treated in discrete intervals)
k	separation along the sequence
s	number of atoms of specific type within a sphere of a fixed radius R

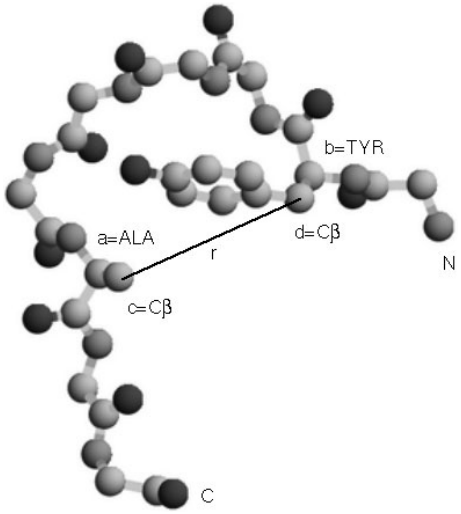


Fig. 11.2. Conformational variables of pair interactions shown on a ball-and-stick model of a protein segment. a, b: amino acids; c, d: atom types; r: spatial distance (treated in discrete intervals); k: separation along the sequence (= 5 here).

Since the frequencies of rare amino acid pairs can be relatively small, procedures to treat sparse data are employed [6].

11.2.2.2 Knowledge-based Potentials for Protein-solvent Interactions

In soluble globular proteins, hydrophilic amino acids tend to be on the exterior of the molecule whereas hydrophobic amino acids are packed in the interior [13]. To quantitatively describe the location of an amino acid in relation to the protein surface, different measures of solvent exposure have been developed. In the present context, the solvent exposure is modeled by the number s of protein atoms that are within a sphere of radius R centered at the position of atom c of amino acid a [5]. If the amino acid is buried in the protein interior, s is large because the surrounding volume is (almost) completely filled by protein atoms. On the other hand, if the amino acid is exposed, part of the volume is occupied by solvent molecules, which results in a smaller s (see Table 11.1 and Figure 11.3). Again, relative frequencies $f^{ac}(s)$ and $f^{c}(s)$ are derived from the database and the net potential for solvent exposure is then

$$\Delta E^{ac}(s) = E^{ac}(s) - E^{c}(s) = -kT \ln\left(\frac{f^{ac}(s)}{f^{c}(s)}\right) \tag{11.5}$$

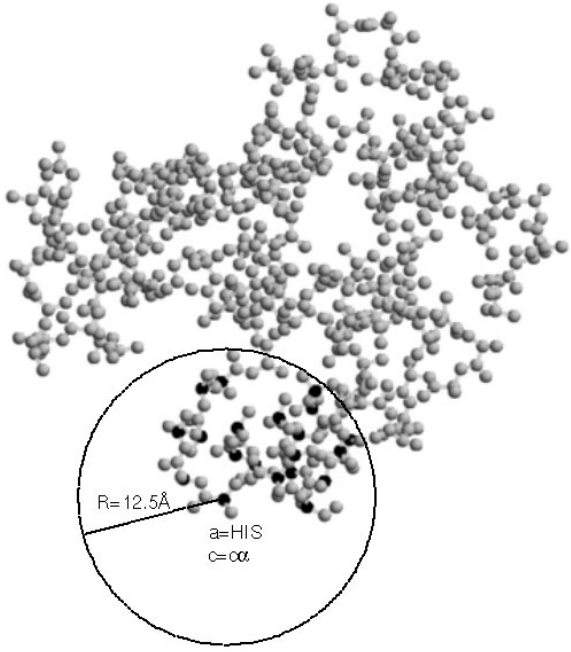


Fig. 11.3. Conformational variables for surface potentials shown on a ball-and-stick model of a protein. a: amino acid; c: atom type; s: number of protein C^{α} atoms (black) within a sphere of radius R. Here, s = 19.

11.2.2.3 Energies of Residues, Combined Energies, Total Energies

The interaction energy of residue i with all other residues of a protein in a particular conformation is obtained by summing over all positions $j \neq i$ in the sequence and over all atom pairs (c, d):

$$\Delta E^{i}_{\text{pair}} = \sum_{j \neq i} \sum_{cd} \Delta E^{a(i)b(j)cdk}(r) \tag{11.6}$$

$a(i)$ and $b(i)$ are the amino acids at i and j, $k = |i - j|$ is the separation of i and j along the sequence, and r is the spatial distance of c and d. The total surface energy for a particular residue is given by

$$\Delta E^{i}_{\text{surf}} = E^{a(i)c}(s) \tag{11.7}$$

The combined energy is a weighted sum of pair and surface terms:

$$\Delta E^{i}_{\text{comb}} = \omega_p \cdot \Delta E^{i}_{\text{pair}} + \omega_s \cdot \Delta E^{i}_{\text{surf}} \tag{11.8}$$

When the energy of residues is plotted as a function of the amino acid sequence position, we obtain an energy profile. High energies in the energy profile point to deviations from the expected energies in native proteins (Figure 11.4).

We refer to the total energy of a protein as the sum over all residue contributions:

$$\Delta E_{\text{comb}} = \sum_{i} \Delta E^{i}_{\text{comb}} \tag{11.9}$$

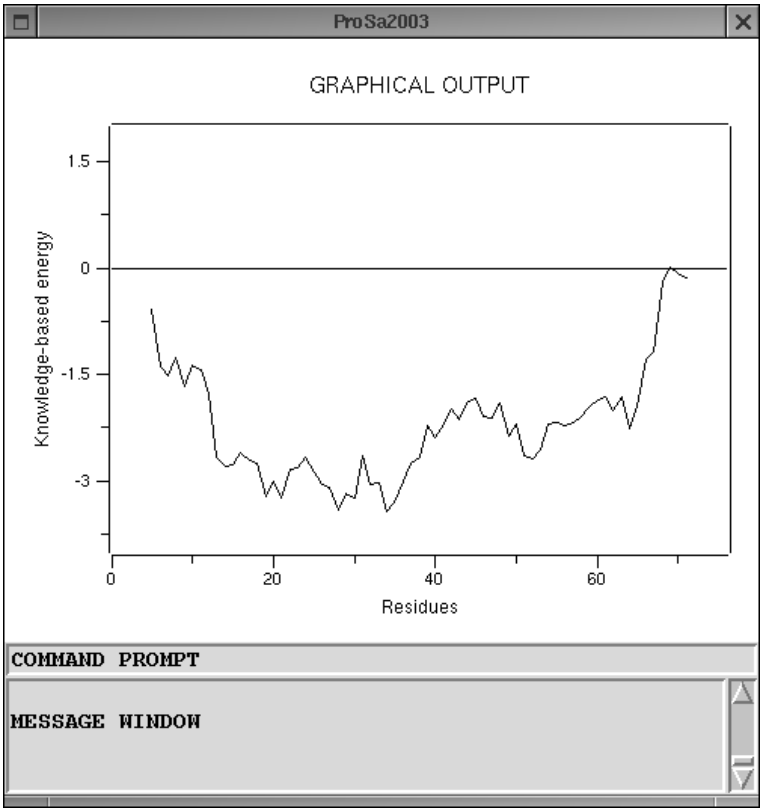


Fig. 11.4. Screenshot of ProSa, displaying an energy profile for a native protein.

11.2.3 Polyprotein, Z-scores

At this point we are able to calculate knowledge-based energies for protein conformations, but we still need to interprete them in a way that helps us in finding out something about the stability of our protein. Our notion of stability depends on the energy of the structure, but it also depends on how this value is related to the energies of all other structures in conformation space. Are there many alternative conformations with an energy higher than that of our structure? Does it significantly depart from the average energy?

To answer these questions, we need to represent conformation space. In our approach it is modelled by using a polyprotein that is constructed from individual proteins by linking them together, ensuring that the linker regions do not violate stereochemical constraints (like Φ and Ψ angles in allowed range, no atom overlaps, and so forth).

The default polyprotein of ProSa consists of 125 protein chains and has a total length of 30681 residues. For a particular protein of length N the polyprotein yields $30681 - N + 1$ distinct conformations and corresponding energies, which are used to compute the z-score according to the following equation:

$$z = \frac{\Delta E_{\text{comb}} - \overline{\Delta E}_{\text{comb}}}{\sigma} \tag{11.10}$$

Z-scores measure the extent to which the energy of a protein structure departs from the mean energy $\overline{\Delta E}_{\text{comb}}$ of all alternative structures on the polyprotein in units of standard deviation σ.

The z-scores of native proteins are usually in a characteristic range (Figure 11.5). Deviations from this indicate non-native properties of the given protein.

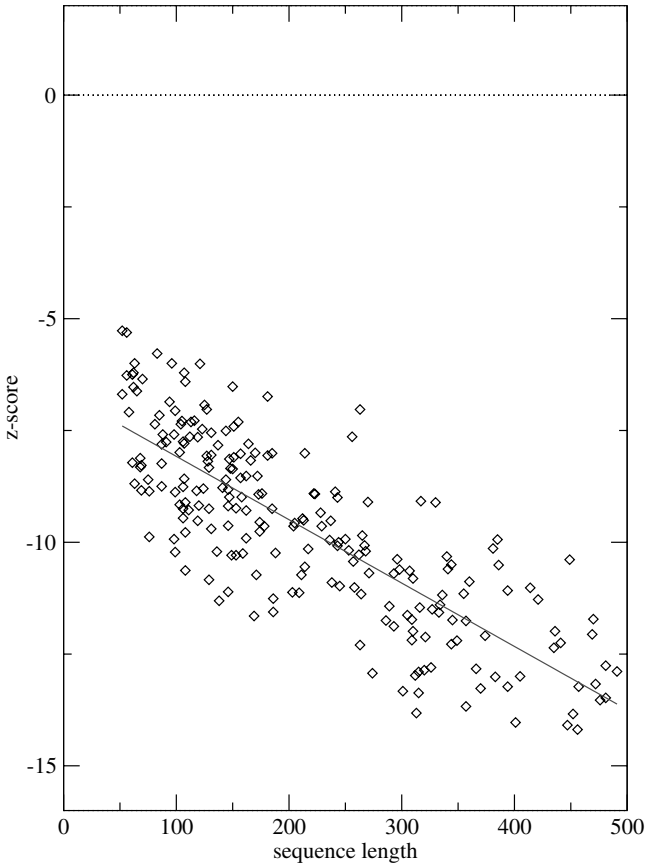


Fig. 11.5. Z-scores of native proteins as function of sequence length.

11.2.4 *In silico* Mutagenesis

In ProSa's simplified description of a protein structure, the protein side chains are represented by their C^β atoms. Since the position of a C^β atom relative to the backbone atoms is the same for all amino acids (except glycine), mutating a residue in ProSa is straightforward: only the amino acid type at the respective sequence position is changed. There is no rearrangement of atoms, nor do we have to construct side chains. The fact that this coarse model is still sufficient to distinguish a native protein structure from a non-native structure rests on a key feature of the knowledge-based potentials employed here: when we record the distribution of C^β–C^β distances in known protein structures, we implicitly capture information on their average environment. Since this environment also includes the side chains, the C^β–C^β potentials strongly depend on the amino acid types involved [6, 12]. Similarly, there is no explicit check for steric clashes that may occur, for example when an amino acid with a short side chain is substituted by an amino acid with a long one. Since the atom distances that result from such substitutions are rarely found in the knowledge base, such clashes are supposed to be penalised by the knowledge-based potentials. Due to these simplifications, computation of the effects of amino acid substitutions is very efficient with ProSa.

For each mutation the difference Δz between the wild type z-score z_{wt} and the mutant z-score z_{mut} is determined by

$$\Delta z = z_{\text{wt}} - z_{\text{mut}} \tag{11.11}$$

We refer to a mutation as *stabilizing* if $\Delta z > 0$, *destabilizing* if $\Delta z < 0$, and *neutral* if $\Delta z = 0$.

11.2.4.1 Single-site Mutability

From these definitions, criteria can be derived that tag certain sequence positions as 'suitable for randomisation'. The criterion used in this work is a high number of nondestabilising mutations. More precisely, we refer to a sequence position p as *randomizable* if there are at least n mutations at p among all possible 20 substitutions that are not destabilizing. The default value of n is 15, which means that at least $\sim$75% of all single-site mutants are required to be not destabilising. For a single-site mutability profile, all 20 point mutations are evaluated for each specified sequence position of a given protein.

11.2.4.2 Multiple-site Mutability

Since an exhaustive evaluation of all possible variants is not feasible if the number of sites exceeds 3 – a group of 4 sites would involve $20^4 = 160000$ z-score calculations – a random sample population of mutants can be generated for specified sets of residues. For each mutant, the amino acids of the selected set are substituted by randomly chosen ones. The number of mutants (default: 10^3) as well as the amino

acids that should be allowed for substitution (default: all 20) can be varied. For the resulting mutants, z-scores and Δz-values are calculated as described above. We then examine how many mutants show a z-score that is at least as low as the wild-type z-score. The higher this ratio, the more likely the selected set corresponds to a protein region where mutations can be introduced without appreciably affecting stability.

11.2.5 Summary

As depicted in Figure 11.1, the main items of ProSa's toolbox are:

1. The *energy profile*, displaying energy values for each single residue in the protein. It allows one to identify energetically unfavorable regions, as well as local effects of amino acid substitutions.
2. The *z-score* and Δz, providing a global estimation of the stability of a wild-type protein and its mutants.
3. The *single-site mutability profile*, showing which residue positions are supposed to be robust towards point mutations.
4. The *multiple-site randomizer*, for generating a population of random mutants for a selected set of residues and estimating multiple-site mutability.

11.3 Protocol

This section describes the individual analysis steps in detail. Each of ProSa's mutagenesis tools is demonstrated with an example. Starting with some general remarks on the usage of the program, we analyze the energy profiles of wild-type and mutant proteins, examine their z-scores, then turn to mutability of single residues, and finally compare two regions of a protein with respect to the acceptance of randomly introduced mutations. Similar to giving a recipe for a wet-lab method, the goal is to supply all information you need to use ProSa for your analysis tasks.

11.3.1 ProSa Setup and Interaction

After installation, ProSa can be executed in several modes (please see the manual for details). The GUI mode provides a simple graphical user interface, consisting of a prompt to enter commands, a message window for program feedback, and a window for graphical output (Figure 11.4). The batch mode allows to run ProSa noninteractively; commands are read from ASCII files and then executed. The possibility of recording the commands you type and storing them in a logfile facilitates the usage of analysis steps you perform repeatedly. All examples discussed below

are included as script files on the CD-ROM. You can easily adapt them with a text editor and make a collection of scripts for your needs.

11.3.2 ProSa Objects

ProSa's data items are called *objects*. An object consists of protein atom coordinates, amino acid sequence, associated energy profiles, specification of potentials, and a set of plot parameters. The number of objects that ProSa can handle at the same time is limited only by memory. A particular object can be accessed by its name, which is created when a protein is loaded or an object is copied.

Copies of objects can be used to display distinct graphs for the same protein. They are also used to generate protein mutants, which are copies of 'wild type' objects with distinct amino acid sequences for the same set of coordinates.

11.3.3 Session 1 (`mut_script1.cmd`)

This session demonstrates how to load a protein structure and substitute some of its amino acids. Energy profiles are generated and used to analyse the local effects of these mutations.

1. Start ProSa – enter at the system command line:

   `prosa2003`

   A window appears. Click the command line.

2. Load the protein – enter:

   `read pdb pdb1ubi.ent wt`

   Load the pdb file `pdb1ubi.ent` and name the resulting object `wt` (wildtype).

3. Perform energy analysis:

   `analyse energy wt`

   Calculate energies for object `wt`.

   `plot`

   Display the energy graph of `wt`. By default, combined energies for all residues are displayed.

4. Edit the graph:

   `winsize wt 10`

   Set the width of a gliding average window to 10.

   `plot`

   The average energy of the first 10 residues is plotted as values for residue 5, then the average of residues 2–11 is plotted as values for residue 6, and so on. This

is used to smooth the often heavily fluctuating graph and get a better overview after problematic regions. As the whole graph lies below the zeroline, nothing seems to be suspicious.

```
draw * wt 1
```

Enables display of pair, surface, and combined energies for `wt`.

```
color comb wt cyan
color surf wt magenta
plot
```

New plots for the various energy terms appear, differently colored for easier distinction.

5. Substitute amino acids:

```
mutate sequence wt 8 E mut1
```

Substitute the amino acid at position 8 with glutamate (E). The resulting object named `mut1` inherits all properties from `wt` except the sequence, its name, and all calculated energies.

```
analyse energy mut1
```

Analyze the energy of the mutant.

```
color * mut1 red
draw * * 0
draw comb * 1
plot
```

Color all mutant energy graphs red, remove all energy graphs, draw combined energies of all objects, display graphs.

An energy difference is visible at the N-terminus, made more clear with a difference plot.

```
diff wt mut1 diff1
```

Calculate difference between object `wt` and object `mut1`. The energy of each residue of `mut1` is subtracted from the energy of the corresponding residue of `wt`. This is done for combined, pair, and surface energy separately. The result is stored in a new object called `diff1` in our example.

```
plot
```

You see graphs for combined energy plus a difference graph.

```
winsize * 1
hide *
show diff1
plot
```

Reset window size (no averaging), hide all objects, and plot only difference graph. The positive peak at residue 8 shows that the energy of the mutant is lower than that of the wild type, indicating a stabilizing mutation.

```
mutate sequence wt 30 E mut2
analyse energy mut2
diff wt mut2 diff2
color * diff2 blue
show diff2
plot
```

The difference plot for the second mutation shows two things: first, the negative peak at residue 30 indicates a destabilizing mutation, and second, the other peaks show that this mutation also affects long-range interactions.

6. Store your results and exit:

   ```
   export plot 1ubi_mutants
   ```

   A postscript file called `1ubi_mutants.ps` is created. It contains the last plot on your screen.

   ```
   print energy *
   ```

   For each object, a file with extension `.ana` is created. It contains the values of pair energy, surface energy, and combined energy for each residue, respectively. It can be used as input for other programs, like plotting tools, spreadsheets, and so forth.

   ```
   exit
   ```

11.3.4 Session 2 (`mut_script2.cmd`)

This session demonstrates how to calculate the z-score of a protein. Four mutants are generated and their z-scores are compared with that of the wild-type structure.

1. Start ProSa and load a protein:

   ```
   prosa2003
   read pdb pdb1ubi.ent wt
   ```

   Load 1ubi and call the resulting object `wt`.

2. Calculate z-score:

   ```
   init zscore
   ```

   Initialize z-score calculation by loading a polyprotein. Without an argument, `pII3.0.short.ply` is loaded.

   ```
   zscore wt
   ```

   Start z-score calculation. The message window indicates the progress of determining the energy distribution on the polyprotein, first for pairwise interactions, then for surface potentials. (The output goes to the command prompt window, which may lie behind the ProSa window.) Since no filename is given to the `zscore` command as second argument, the out goes to the screen.

   See Figure 11.6 for an explanation of the displayed values.

```
Hide & Seek on polyprotein pII3.0.short.ply - selection of parameters
molecule     seq-l  zp-comb  zp-pair  zp-surf  rk-comb  rk-pair  rk-surf
wildtype        76    -8.94    -5.74    -6.48        1        1        1


             z1-comb  z1-pair  z1-surf  ep-comb  ep-pair  ep-surf
               -4.58    -3.94    -4.49  -101.52   -47.82   -10.74


             em-comb  em-pair  em-surf  es-comb  es-pair  es-surf
               68.65    12.22    11.29    19.04    10.47     3.40
```

Fig. 11.6. Output of **zscore** command (lines are wrapped).

`seq-l` Sequence length of object `wildtype`.

`zp-comb, zp-pair, zp-surf` Z-scores of object `wildtype`. There is one value for pair potentials, one for surface potentials and one for the combination of them. With respect to the sequence length, all z-scores indicate that our protein is native-like (Figure 11.5).

`rk-comb, rk-pair, rk-surf` Rank (relative position) of object `wildtype` in an energy sorted list of all polyprotein fragments with length `seq-l`. A rank of 3 would tell us that there are 2 conformations with the sequence of `wildtype` that have a lower energy than our protein. A rank of 1 means that we did not find any conformation that fits better to our sequence than the one of `wildtype` (at least not in our sample of conformation space). This again is an indication for a native-like protein.

`z1-comb, z1-pair, z1-surf` Z-scores of fragment of lowest energy found in the polyprotein.

`ep-comb, ep-pair, ep-surf` Energy of object `wildtype`, for combined, pair and surface potentials, respectively.

`em-comb, em-pair, em-surf` Average energy values of all fragments derived from the polyprotein.

`es-comb, es-pair, es-surf` Standard deviation of energies derived from the polyprotein.

3. Substitute amino acids:

   `mutate sequence wt 46 P mutant1`

   Substitute the amino acid at position 46 of object `wt` with a proline (P).

   `mutate sequence wt 5 E mutant2`
   `mutate sequence mutant2 32 P mutant3`
   `mutate sequence mutant3 47 L mutant4`

   Substitute three other amino acids. To accumulate mutations, repeatedly apply `mutate sequence` to mutant objects.

   `list objects`

   A list of all currently existing objects is displayed. The mutation is shown as a suffix to the protein name.

   `zscore *`

   We use the `*` to apply the `zscore` command to all objects we have generated so far. The result (Table 11.2) shows that `mutant1` has a higher z-score than

`wt`, whereas `mutant2-mutant4` have a lower one. Hence, A46P is supposed to be a stabilizing mutation, whereas the others are destabilizing (also note the ranks of `mutant4`!).

4. Exit ProSa: **exit**

Table 11.2. Example of z-scores of four different ubiquitin mutants.

object	molecule	seq-l	zp-comb	zp-pair	zp-surf	rk-comb	rk-pair	rk-surf
wt	wt	76	−8.94	−5.74	−6.48	1	1	1
mutant1	wt_A46P	76	−9.54	−6.64	−6.63	1	1	1
mutant2	wt_V5E	76	−7.90	−4.86	−5.86	1	1	1
mutant3	wt_V5E_D32P	76	−7.68	−4.46	−5.84	1	1	1
mutant4	wt_V5E_D32P_G47L	76	−6.71	−3.92	−5.08	1	3	1

11.3.5 Session 3 (`mut_script3.cmd`)

Now we demonstrate how to examine the mutability of specified residues in a protein. Two positions in an immunoglobulin structure are compared with respect to the number of stabilizing mutations among all possible substitutions. A single-site mutability profile is generated for two regions of the structure, pointing to sites that may be more important for structural stability than others.

1. Start ProSa, initialise a z-score calculation, and load a protein:

   ```
   prosa2003
   init zscore
   read pdb pdb1aqk.ent,H, Fd
   ```

   Load chain H of 1aqk and name the resulting object `Fd`. This is the structure of an immunoglobulin Fd fragment.

2. Substitute amino acids exhaustively:

   ```
   mutate sequence Fd 69 *
   ```

   By using `*` for the amino acid that we want to introduce at position 69, we generate all 20 mutants at once. Since we did not specify an object name, the mutants will be called `Fd_x69y`, where x is the wild type amino acid at position 69 and y is the introduced amino acid. A subsequent `zscore *` would calculate all z-scores for the Fd wildtype and its mutants. The next command facilitates this kind of analysis by doing z-score calculations *and* subsequent comparison with the wild type:

   ```
   analyse mutability Fd 69
   ```

   First, this command generates all mutants (Fd_I69y) if they do not already exist.

Then z-scores for all mutant objects are calculated and the portion of stabilizing, destabilizing, and neutral mutations is recorded. These values are finally sent to the screen or to a file, if a filename was given as additional argument.

`analyse mutability Fd 9`

Checking all substitutions at position 9, we see that this position is much less sensitive to amino acid exchanges.

3. Calculate mutability profile:

   `analyse mutability Fd 98-106,205-210 Fd_xmpl`

   To check single-site mutability for more than one residue, we specify a list of positions. Given a computing time of roughly 20 s per z-score calculation for a protein of this length on a 1.4 GHz cpu, the analysis takes approximately 1.75 h (15 residues $\times$ 20 amino acids $\times$ 20 secs.). The output is collected in four files: `Fd_xmpl.slp` holds the z-scores for wild type and mutants and `Fd_xmpl.mut_comb`, `Fd_xmpl.mut_pair`, and `Fd_xmpl.mut_surf` hold the mutability profiles for the three energy terms, respectively.

   `exit`

   As can be seen in Figure 11.7, all randomizable residues are marked with a `+`. The residues of the first set (98–106) highly tolerate single-site mutations. This result is promising, since the residues lie within the hypervariable region of the Fd fragment. The residues of the second set (205–210) are part of a β-sheet in the C-terminal half of the Fd fragment. Each of them is much more restricted in terms of allowed amino acid substitutions. Remarkably, only a cysteine is accepted at position 205 (bridging to a cysteine at 149); all other mutations are regarded as destabilizing.

```
Single-site mutability for Fd (zp-comb)
pos  aa_wt  stabilising  destabilising  neutral  >= rnd_cutoff[15]
-------------------------------------------------------------------
98       V          18             1        1            +
99       L          19             0        1            +
100      F          16             3        1            +
101      Q           6            13        1
102      Q           7            12        1
103      L          18             1        1            +
104      V          16             3        1            +
105      L          18             1        1            +
106      Y          16             3        1            +
205      C           0            19        1
206      N           8            11        1
207      V           0            19        1
208      N           7            12        1
209      H           9            10        1
210      K           2            17        1
```

Fig. 11.7. Single-site mutability profile for residues 98–106 and 205–210 of PDB file 1aqk, chain H (Fd fragment).

11.3.6 Session 4 (`mut_script4.cmd`)

The final session demonstrates how to characterize a protein region as randomizable. For a set of solvent-exposed residues of an immunoglobulin structure, 10^3 mutants are randomly generated. We examine how many of these mutants are destabilized with respect to the wild type. The analysis is repeated with an alternative set of residues that correspond to part of the natural epitope of the immunoglobulin structure.

1. Start ProSa and load a protein:
   `prosa2003`
   `read pdb pdb1aqk.ent,H, Fd`
   Load chain H of 1aqk (immunoglobulin Fd fragment) and name the resulting object `Fd`.

2. Set number of mutants:
   `nr_mutants = 1000`
   The variable `nr_mutants` determines the size of the sample population, 1000 is the default value.

3. Create and analyse pool of mutants:
   `init zscore`
   `randomise sequence Fd 153,180,182,184 Fd_epi1`
   This command first derives 10^3 mutants from the object `Fd` by substituting all amino acids in the specified positions with randomly chosen ones. For each mutant, z-scores are calculated and compared to the `Fd` z-scores. Since 1001 z-scores have to be determined, this calculation lasts about 5.5 h. Results are written to two files: `Fd_epi1.slp` contains all the z-scores, and `Fd_epi1.nrm` holds a summary of the number of stabilized, destabilized, and neutral mutants within the sample population (Figure 11.8).
   Only a very small number of mutants (3.6%) shows a combined z-score that is wild type-like or below, the majority is destabilised.
   Now repeat the experiment with a different region, namely a part of the natural epitope of the immunoglobulin:

   `randomise sequence Fd 100,104-106 Fd_epi2`
   Here, the situation is almost reversed: only a minority of the mutants (1.1%) exhibit a combined z-score that indicates destabilization (Figure 11.9 and 11.10 for a comparison of the z-score distributions). This result is in good agreement with our prior knowledge about immunoglobulin structures: the amino acids of the epitope are supposed to be primarily selected for binding the antigen, rather than for their contribution to structural stability.

```
Randomisation of Fd, residue(s) 153,180,182,184
Total number of mutants analysed: 1000

zp-comb of wildtype: -7.78
nr_stabilised     nr_neutral     nr_destabilised
-------------------------------------------------
    36 (3.6%)      0 (0.0%)        964 (96.4%)


zp-pair of wildtype: -6.22
nr_stabilised     nr_neutral     nr_destabilised
-------------------------------------------------
    25 (2.5%)      0 (0.0%)        975 (97.5%)


zp-surf of wildtype: -5.41
nr_stabilised     nr_neutral     nr_destabilised
-------------------------------------------------
    95 (9.5%)      0 (0.0%)        905 (90.5%)
```

Fig. 11.8. Result of randomization on a region of an Fd fragment (PDB file 1aqk, chain H). The region consists of four spatially close residues in the C-terminal half of the chain.

```
Randomisation of Fd, residue(s) 100,104-106
Total number of mutants analysed: 1000

zp-comb of wildtype: -7.78
nr_stabilised     nr_neutral     nr_destabilised
-------------------------------------------------
   989 (98.9%)      0 (0.0%)         11 (1.1%)


zp-pair of wildtype: -6.22
nr_stabilised     nr_neutral     nr_destabilised
-------------------------------------------------
   977 (97.7%)      0 (0.0%)         23 (2.3%)


zp-surf of wildtype: -5.41
nr_stabilised     nr_neutral     nr_destabilised
-------------------------------------------------
   989 (98.9%)      0 (0.0%)         11 (1.1%)
```

Fig. 11.9. Result of randomization on part of the antigen-binding region of an Fd fragment (PDB file 1aqk, chain H).

11.3.7 Tips & Tricks

- If you set the ProSa variable `log` to 1, all the instructions you type are recorded and stored in the file `ProSa2003.log`. This file can then be used as a template for similar analysis tasks. Simply make the necessary changes with a text editor and execute it as a batch job (see the manual for details). A special example of a command file is the `prosa-startup` file: it is run automatically each time ProSa is started. By adapting this file you can customize your sessions and initialize ProSa with the parameters you found to be optimal for your tasks.

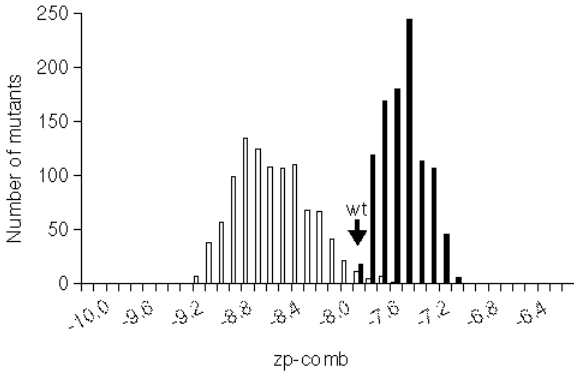


Fig. 11.10. Z-score distribution for two randomized regions of an Fd fragment (PDB file 1aqk, chain H). White bars: residues 100, 104–106 (part of antigen binding region), black bars: residues 153, 180, 182, 184 (in loop of C-terminal half of the chain). The arrow marks the z-score of the wild type.

- You can store your laboriously generated mutants by saving them as binary backbone files (see command `write bbn`). The files are created in your current working directory. To restore them, use the command `read bbn`. You can organize your *.bbn files in several folders and use the variable `bbn_dir` to tell ProSa where they are located.
- Pairwise interactions between amino acids are calculated in a distance range from 0 to 15 Å. You can adjust this range if you want to neglect certain interactions. For example if you want to mask any energy contribution of close contacts, set the variable `pot_lb` to 4. The energy of pair interactions between 0 and 4 Å is then zero. Similarly, you can focus on interactions for residue pairs with a certain sequence separation. For example, if you are interested only in short-range energy contributions (e. g., sequence separation $k \leq 9$), set `lower_k` to 1, `upper_k` to 9 (default: 600).
- Z-score analysis leads to a lot of result values (Figure 11.6). In general, *zp-comb*, *zp-pair*, and *zp-surf* together with the respective ranks (which generally should be 1 for native-like structures) are sufficient for judging the result.
- ProSa includes C^{α} and C^{β} potentials. In view of our previous observations, it does not seem to be necessary to use both for mutation analysis. Δz-values correlate highly for C^{α}, C^{β}, and C^{α}+C^{β} potentials.
- For the combination of pair and surface potentials these terms must be properly weighted. The default values (`factor_pair=1`, `factor_surf=5`) might be inappropriate for small proteins (~60–80 residues). For these proteins, smaller values for the weight of the surface energies are recommended (`factor_pair=1`, `factor_surf~3`). By setting `combine type sdev`, the standard deviations of pair and surface energy distributions are used for determining the weights, which in general yields appropriate values.

- You can adjust the minimum number n of nondestabilizing mutations that are required to characterize a sequence position as randomizeable. For example, if you want to be very restrictive and regard a position as randomizeable only if all possible amino acid exchanges are either stabilizing or neutral, set `rnd_cutoff` to 20.
- By default, all 20 amino acids are used for exhaustive or random substitutions. If you want to restrict the amino acid alphabet, set the variable `alphabet` to a string like `FWYH`. In this example, only aromatic amino acids will be used. Another application is to prevent mutants that are to be expressed in *Escherichia coli* from having cysteines.
- The *.slp-file resulting from the `randomise sequence` command holds the names of all mutants in the sample population. Since these names contain the randomized positions as well as the substituted amino acids, they can be used to derive the distribution of amino acids for certain positions. By analyzing this distribution for the subset of nondestabilized mutants, it is possible to gain hints about amino acid preferences.
- Finally, if you are not sure about the syntax of a command or the name of a variable, make use of the `help` command. Without argument, it lists all commands and variables. With a string as argument, it shows you information on all the commands and variables that contain this string.

11.4 Troubleshooting

You need to be careful about several things when performing *in silico* mutagenesis with ProSa. In contrast to many wet-lab setups, where you have to calibrate a lot of parameters and prepare the right environment for a successful realisation, the program leads to an outcome in a rather short time. Moreover, you may get *some* result even with unreasonable input. This includes the danger that less care is taken about the ingredients for the experiment (input data), the general conditions (program parameters), and the interpretation of the results (output data). In this section we address the major pitfalls:

- In general, you should check the quality of the wild-type structure model before you estimate the effects of mutations. The results may be misleading if your wild-type structure derives from a homology modeling study rather than from an X-ray model with 2.0 Å resolution. Also be aware of the context of your wild-type structure: Has it been crystallized in a complex with another protein chain? Are there any ligands that might influence interactions afflicted by the intended mutation? A mutation might be neutral if introduced when the protein is complexed, but destabilizing for the uncomplexed chain.
- If you work with a model that consists of only a C^{α} trace of the protein backbone, make sure you only use C^{α} potentials, other types could lead to a program crash.

- An issue that has already been mentioned in section 11.2 is residue numbering. Residues listed in the ATOM records of a PDB file are not necessarily numbered consecutively, may not start with residue number 1, may have an additional insertion code, and may even be negative (see [2] for details). ProSa, in contrast, uses a sequential index to address residues: the first residue gets number 1, the second residue number 2, and so on. In addition, some of the residues present in a PDB file may be skipped by ProSa, for example, because they are not a standard amino acid. As a consequence, you have to make sure that the residue(s) you substitute with `mutate sequence` or `randomise sequence` really correspond to those you have in focus. One hint is the correct wild-type amino acid, as displayed in the object name of the mutant. To get a list of PDB residue numbers of a certain object and how they are mapped to the sequential index, use `print residue mapping {object}`.
- Consecutive use of the mutagenesis commands can quickly lead to a huge number of objects. You should clear memory from time to time by using the `delete` command.
- ProSa potentials derive from a knowledge base of globular, soluble protein structures. If they are applied to the analysis of proteins with different characteristics (such as membrane proteins), the results may lead you astray.

11.5 Major Applications

In the past decade ProSa has been used to address a variety of problems in protein structure research. A list of references to the relevant publications is provided at the ProSa website [14]. Originally the program was designed to spot errors and faulty parts in protein structures – whether the structures were determined by experiment or by modeling does not matter [9].

In the present article, additional features of ProSa for performing *in silico* mutagenesis are described. They extend its usage from the evaluation of *structure models* to the evaluation of *sequence models*. The applicability of ProSa in this context has been demonstrated by Babajide et al. [15, 16], who used ProSa to study neutral networks in protein evolution. In a joint project with a wet-lab group, ProSa supported the identification of epitopes on protein scaffolds, that is, regions on the protein surface that can be used for the design of new binding properties without destabilizing the structure. It is encouraging that the experimental results corroborated the epitopes that were suggested with the help of ProSa (Fiedler et al., unpublished).

The constant refinement of techniques for directed protein evolution also involves the development of increasingly sophisticated *in silico* tools. This co-evolution of experimental and computational methods enriches our toolkit for finding the sequence that fits. It is this mutual impact which makes ProSa a valuable component in designing both experiments and proteins.

References

1. M.J. Sippl. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, *5*:229–235, **1995**.
2. RCSB. *PDB Format Description Version 2.2*. http://www.rcsb.org/pdb/docs/format/pdbguide2.2/guide2.2_frame.html
3. C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, *181*:223–230, **1973**.
4. C.J. Epstein, R.F. Goldberger, and C.B. Anfinsen. The genetic control of tertiary protein structure: studies with model systems. In: *Cold Spring Harbor Symposium on Quantitative Biology*, **1963**.
5. M.J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding: an approach to the computational determination of protein structures. *J. Comput. Aided Mol. Des.*, *7*:473–501, **1993**.
6. M.J. Sippl. Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, *213*:859–883, **1990**.
7. S. Vajda, M. Sippl, and J. Novotny. Empirical potentials and functions for protein folding and binding. *Curr. Opin. Struct. Biol.*, *7*:222–228, **1997**.
8. W.A. Koppensteiner and M.J. Sippl. Knowledge-based potentials-back to the roots. *Biochemistry (Mosc)*, *63*:247–252, **1998**.
9. M.J. Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, *17*:355–362, **1993**.
10. M.J. Sippl and M. Jaritz. Predictive power of mean force potentials. In: H. Bohr and S. Brunak, eds., *Protein Structure by Distance Analysis*, 113–134. IOS Press, **1994**.
11. M.J. Sippl, M. Jaritz, M. Hendlich, M. Ortner, and P. Lackner. Applications of knowledge based mean fields in the determination of protein structures. In: Doniach, ed., *Statistical Mechanics, Protein Structure and Protein–Substrate Interactions*, 297–315. Plenum Press, **1994**.
12. M.J. Sippl, S. Weitckus, and H. Flockner. In search of protein folds. In: Merz and LeGrand, eds., *The Protein Folding Problem and Tertiary Structure Prediction*, 353–407. Birkhaeuser, **1994**.
13. W. Kauzmann. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.*, *14*:1–63, **1959**.
14. Center of Applied Molecular Engineering. Survey of literature that references ProsaII. http://lore.came.sbg.ac.at: 8080/CAME/CAME_EXTERN/PROSA/references_html
15. A. Babajide, I.L. Hofacker, M.J. Sippl, and P.F. Stadler. Neutral networks in protein space – a computational study based on knowledge-based potentials of mean force. *Folding Design*, *2*:261–269, **1997**.
16. A. Babajide, R. Farber, I.L. Hofacker, J. Inman, A.S. Lapedes, and P.F. Stadler. Exploring protein-sequence space using knowledge-based potentials. *J. Theo. Biol.*, *212*:35–46, **2001**.

12 RNA Folding *in silico*

Christoph Flamm, Ivo L. Hofacker, Peter F. Stadler

12.1 Introduction

Bioinformatics has invaded most 'wet-lab' environments, where databank searches in `Genbank` and `PubMed`, and thus algorithms such as `blast`, have become indispensable tools in the daily routine. This is largely due to the fact that the most often used services are available as convenient web applications and require little or no thought about such details as getting and installing software or reading technical manuals.

Many tools that could prove to be extremely useful for your *particular* field of research are being developed in bioinformatics groups all over the world. These programs are typically just a couple of mouse clicks away and can be downloaded and used free of charge. In most cases, however, these tools are rarely converted into webtools and usually don't come with a fancy graphical user interface.

In RNA bioinformatics, only a few basic algorithms, namely those for structure prediction based on thermodynamic rules, are available as web tools. In addition, RNA structure prediction is a computationally rather demanding process, so that the sequence lengths that can be dealt with on the web are limited. Because of these limitations we recommend that you install the software locally on a computer in your lab (or your laptop).

Functional RNA molecules, whether natural or produced in the lab through directed evolution, typically require distinctive secondary structures to fulfill their function; for a nice example we refer to Schwienhorst [8]. These structures serve as a scaffold that allows the formation of, e. g., a catalytic site. Thus, sequence constraints observed in RNA molecules selected for a particular function, such as binding or catalysis, may be due to direct involvement in that function or due to stabilization of the structure. Predicted RNA secondary structures can be most helpful to identify such structural constraints and to interpret the results of a directed evolution experiment in terms of structure–function relationships.

Before we start with the recipes, let us very briefly give a few hints on the background of RNA folding algorithms used below. RNA secondary structures are described as graphs (Figure 12.1), so that entire nucleotides are thought of as nodes and the backbone of the molecule and the Watson–Crick and GU base pairs are represented as edges. The energy of a particular sequence in a given secondary structure is approximated as the sum of contributions from the 'loops' (i. e,

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst
Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-30799-0

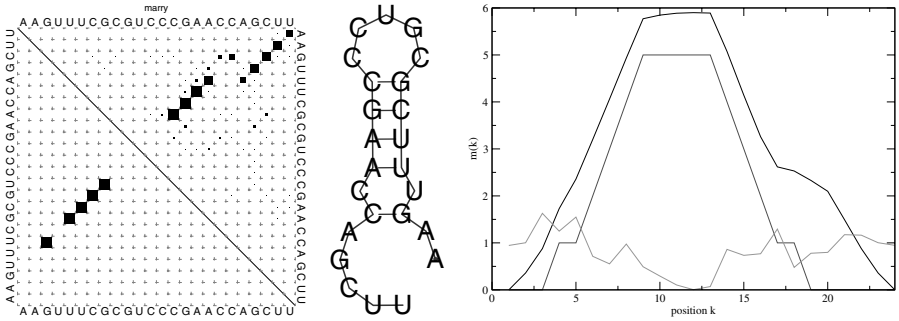


Fig. 12.1. Dot plot, secondary structure graph, and mountain plot for the example sequence `marry`.

the faces of the drawing in the plane). In this view, stacked base pairs are interpreted as a special type of loop. Energy parameters for the loops depending on their type (stacked base pairs, hairpin loop, interior loop, bulge, or multi-branched loop) and sequence have been compiled over the last 20 years based on melting experiments with a huge number of small molecules [6]. The RNA folding algorithms discussed below use a so-called dynamic programming approach to find, e. g., the secondary structure that minimizes the energy, given a particular sequence. The results of the computations are therefore *exact within the energy model*. The energy model, however, is based on (1) the *assumption* that loop contributions are additive, and (2) on experimental data that are of course affected by all the usual sources of noise. For a recent review on RNA structure prediction see e. g., [2, 10].

12.2 Materials

If you just want to produce the occasional drawing of an RNA secondary structure you may skip to Section 12.2.2.

If you choose to do your bioinformatics computations locally we strongly recommend that you set up a `Linux workstation` for this purpose. Why? (1) Because it is a lot cheaper than the alternatives. (2) Because, after an initial phase of learning how to do it, it is much more efficient, in particular if you have to work with a large number of sequences on a regular basis. (3) Because the overwhelming part of the more specialized software (the things that you can't get as webservices) are developed in a UNIX environment. (4) Because you can use an outdated PC so even hardware costs are not an issue. A Pentium I with 200 MHz, 64 Mbyte, and 3 Gbyte disk space is easily sufficient for almost all computations that you might want to do with RNAs smaller than, say, 23S RNA.

Here is the shopping list:

1. Buy any recent Linux distributions (e. g., from Redhat, Suse, etc.) for a few Euros (or $) in your university bookstore. Alternatively, your computer center might support Linux so that you can just install it from the web or borrow CDs from them for free.
2. Make sure that your system has the following basic packages installed: A compiler for the C programming language (such as gcc), the Perl scripting language, and tcl/tk. Currently, all common distributions include these features in their standard setup.
3. Once you have a standard Linux installation, you need to download and install some basic bioinformatics software. For the examples here, all you need are listed below

Table 12.1. Basic Bioinformatics Software

Name	URL
Vienna RNA Package	http://www.tbi.univie.ac.at/~ivo/RNA/
ClustalW/ClustalX	http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/
Grace (xmgrace)	http://plasma-gate.weizmann.ac.il/Grace/

Hints:

- If you want to use a cheap laptop for bioinformatics work, make sure beforehand that Linux runs on it without problems. Some products have serious problems because the plug-and-play features of the BIOS cannot be disabled. Usually, information on such problems can be found on the web.
- **Mac users:** The new Mac OS X has a Unix-style operating system underneath the stylish desktop, hence you can work on it just as on a Linux box. However, most of the development environment is not installed by default. The Installation Guide on the CD contains information on what to do.

12.2.1 Typographical Conventions

- Constant-width font is used for command names, variable names and other literal text like input and output in the terminal window.
- Lines starting with a $ within a literal text block are commands. You should type the text following the $ into your terminal window, finishing by hitting the return-key. (The $ signifies the command line prompt, which may be different on your system).
- All other lines within a literal text block are the output from the command you just typed.

As an example for an installation we describe the installation of the Vienna RNA Package. Full instructions are included in the installation instructions on the

CD. Point your web browser to the URL given in Table 12.1 above and download the source code of the latest version of the `Vienna RNA Package` [5].

1. Unpack the tar file by running:
   ```
   $ gunzip ViennaRNA-1.5.tar.gz
   $ tar -xvf ViennaRNA-1.5.tar
   ```
2. To configure, build, and install the package just run:
   ```
   $ cd ViennaRNA-1.5
   $ ./configure --with-cluster
   $ make all
   $ make install
   ```

To run the last command, which installs the main programs of the `Vienna RNA Package` into the default location (`/usr/local/bin/`), you need superuser (`root`) privileges. In addition, a couple of scripts and example programs are installed in the directory `/usr/local/share/ViennaRNA/bin/`. The `--with-cluster` option compiles a few extra programs for cluster analysis, which we will use later.

The installation location can be controlled through options to the `configure` script. For example, to change the default installation location to the directory `FOO` in your home directory use:

```
$ ./configure --prefix=$HOME/FOO
```

Have a look at the file `INSTALL`, distributed with the `Vienna RNA Package`, for more detail or read documentation on the web. Wherever you install the main programs of the `Vienna RNA Package`, make sure the path to the executables shows up in your `PATH` environment variable. Similarly, the `MANPATH` environment variable contains the list of directories searched for man pages (online help). To check the contents of, e. g., the `PATH` environment variable, run:

```
$ echo $PATH
```

12.2.2 RNA Web Services

If a few simple structure predictions is all you want to do, if you are baffled by the command line, and compiling programs just seems too technical, then web services are for you.

Several useful sites for doing RNA structure analysis are available on the web. Most of the tasks described below can be performed with the Vienna RNA web server [3] at `http://rna.tbi.univie.ac.at/`. Currently, it offers web access to the `RNAfold`, `RNAalifold`, and `RNAinverse` programs whose command line usage is shown below.

Another excellent site for RNA structure prediction is Zuker's `mfold` server at `http://www.bioinfo.rpi.edu/~zukerm/rna/`.

Web servers are also a good starting point for novice users since they generally provide a more intuitive interface. Moreover, the Vienna RNA server returns the

equivalent command line invocation for each request, making the transition from web services to locally installed software easier.

Web servers are not ideal for analyzing many or very long sequences. Command line tools, on the other hand, are ideally suited for automating repetitive tasks. They can even be combined in pipes to process the results of one program with another.

12.3 Protocols

12.3.1 Secondary Structures for Individual Sequences

The two programs to compute secondary structures from a single sequence are called `RNAfold` and `RNAsubopt`. Both programs read input from `stdin` and write output to `stdout`. In the basic mode `RNAfold` returns only a single optimal MFE (minimum free energy) structure, but `RNAsubopt` [9] generates a whole list of suboptimal structures for a predefined energy interval above the MFE structure. `RNAfold` computes additional summary information of the structure ensemble at thermodynamic equilibrium if used with the `-p` option.

The input file format for both programs is fairly simple. An input file contains one or more sequences. A sequence must appear as a single line in the file without embedded white spaces. A sequence may be preceded by a special line starting with the '`>`' character followed by a sequence name. This name is used by `RNAfold` as the root name for the postscript output files for this sequence.

1. Prepare a sequence file for input:

   ```
   $ echo "> marry\nAAGUUUCGCGUCCCGAACCAGCUU" > marry.seq
   $
   ```

   You can check the content of the file `marry.seq` using the command `cat` or `more`:

   ```
   $ cat marry.seq
   > marry
   AAGUUUCGCGUCCCGAACCAGCUU
   $
   ```

   Of course you can create the input file with your favorite text editor, such as `emacs` or `vi`. Windows and Mac users beware: if you use a word processor (such as `MSWord`) you *must* save the file as plain text, not in the word processor's native file format. (Otherwise your file will be garbled by the word processor's formatting information!)

2. Compute a single optimal structure:

   ```
   $ RNAfold < marry.seq
   > marry
   ```

```
AAGUUUCGCGUCCCGAACCAGCUU
..(.((((.....)))).)..... ( -1.60)
$
```

The last line of the text output contains the predicted MFE structure in bracket notation and its free energy in kcal mol^{-1}. A dot in the bracket notation represents an unpaired position, and a base pair (i, j) is represented by a pair of matching parentheses at positions i and j.

In addition to the text output, a postscript file is produced, which is a high quality drawing of the secondary structure (Figure 12.1). Since a line with the sequence name was included in the input file `marry.seq` the postscript file is named `marry_ss.ps` instead of the default name `rna.ps`. Postscript files can be printed on any postscript-capable printer or viewed onscreen using a postscript viewer such as `gv` or `gsview`.

3. Compute the MFE structure and additional equilibrium ensemble properties:

```
$ RNAfold -p < marry.seq
> marry
AAGUUUCGCGUCCCGAACCAGCUU
..(.((((.....)))).)..... ( -1.60)
,{{{{{(.....)}}}...}}}, [ -2.71]
 frequency of mfe structure in ensemble 0.165493
$
```

The last two lines are new, compared to the text output without the `-p` option and are a rough measure for the well-definedness of the MFE structure. The line before the last line shows a condensed representation of the pair probabilities, similar to the bracket notation, followed by the ensemble free energy in kcal mol^{-1}. The structure string contains additional symbols coding for the pairing tendency of that position.

On the last line, the frequency of the MFE structure in the equilibrium ensemble is given. An MFE structure is well-defined if the frequency within the equilibrium ensemble is high and the two structure strings look similar.

Besides the text output the postscript file `marry_dp.ps` is generated (if the input file contained no line with the sequence name the filename defaults to `dot.ps`). The 'dot plot' shows the pair probabilities within the equilibrium ensemble as an $n \times n$ matrix and is an excellent way to visualize structural alternatives (Figure 12.1). A square at row i and column j indicates a base pair. The area of a square in the upper right half of the matrix is proportional to the probability of the base pair (i, j) within the equilibrium ensemble. The lower left half shows all pairs belonging to the mfe structure. The matrix of a sequence with well-defined MFE structure should show only a few very small additional squares in the upper right half compared to the lower left half.

4. Transforming a dot plot into a mountain plot:

```
$ mountain.pl marry_dp.ps | xmgrace -pipe
```

A mountain plot is a structure representation that works fine even for very long sequences, for which secondary structure graphs and dot plots become cumbersome. It is a xy diagram plotting the number of base pairs enclosing a sequence position (or for pair probabilities, the *average* number of enclosing base pairs) along the ordinate *versus* the sequence position along the abscissa. The `Perl` script `mountain.pl` transforms a dot plot into the mountain plot coordinates, which are printed to `stdout`. You can visualize the output from `mountain.pl` with any xy plotting program, e. g. `xmgrace`.

The resulting plot shows three curves: two mountain plots derived from the MFE structure and the pairing probabilities and a positional entropy curve. Well-defined regions are identified by low entropy. By superimposing several mountain plots, structures can easily be compared.

5. Suboptimal folding:

```
$ RNAsubopt -e 1 -s < marry.seq
> marry [100]
AAGUUUCGCGUCCCGAACCAGCUU   -160    100
(((((((((.....))))...))))) -1.60
..(.((((.....)))).).....  -1.60
(((((((((.....)))...)))))) -1.30
.(((((((.....))))...))).  -1.10
....((((.....))))).......  -1.00
.(((((((.....)))...))))).  -0.80
$
```

The text output shows an energy-sorted list (option `-s`) of all secondary structures within an energy interval of 1 $\mathrm{kcal\,mol^{-1}}$ of the MFE structure. The sequence example shows a degenerate ground state: only one of the two possible MFE structures is returned by `RNAfold`. The energy interval can be controlled by the argument passed to the option `-e`.

The number of suboptimal structures grows exponentially with sequence length and therefore this approach is tractable only for sequences with fewer than 100 nt. To keep the number of suboptimal structures manageable, the option `-noLP` can be used, forcing `RNAsubopt` to produce only structures without isolated base pairs.

12.3.2 Consensus Structures of a Sample of Sequences

In nature, to preserve function a strong selective pressure may act on the secondary structure of functional RNA molecules, while the sequences diverge. This effect

makes it possible to use computer programs to infer the conserved structure from sequence covariation.

`RNAalifold` [4] combines the standard energy model for RNA folding, as used, for example, by `RNAfold`, with a covariance term, leading to a higher accuracy of the predicted consensus structure as compared to the prediction from a single sequence. The program is used much the same way as `RNAfold`, with the exception that it uses a sequence alignment as input instead of a single sequence.

As an example, we use RNAI, an antisense repressor of the replication of some *Escherichia coli* plasmids with a ColE1 origin of replication.

1. Prepare an input sequence alignment:

```
$ clustalw RNAI.seqs > RNAI.out
```

`RNAalifold` uses a multiple sequence alignment in `Clustal` format as input.

2. Compute the consensus structure from the alignment `RNAI.aln`:

```
$ RNAalifold -p RNAI.aln
10 sequences; length of alignment 108.
CGUAUUUGGUGGCUGCGCUCUUCUACAGCCAG_UUACCACGGUUCAAAAAUUUGCCAGC_UUAGUGAACCUUCGAAAAA_CCACC__UGCCAG_GGUGGUUUUUUCGU
.......(((..(((.(((.......))))))...)))..(((((((.((.........)).)))))))..(((((((.((((((.......)))))).))))))). (-26.87)
.......(((..(((.(((.......))))))...)))..(((((((.{{.........}}.)))))))..(((((((.((((((.......)))))).))))))). [-27.98]
 frequency of mfe structure in ensemble 0.164884
```

Just like `RNAfold`, `RNAalifold` writes two files in postscript format, namely `alirna.ps` and `alidot.ps`, which display the consensus secondary structure and the dotplot (Figure 12.2). The dot plot uses color to convey information on sequence variation. The color encodes the number of different base pairs observed and ranges from red (conserved pair type) to blue, where all 6 pair types occur. Unsaturated colors make pairs that cannot be formed by all sequences.

12.3.3 Sequence Design

Occasionally, you may encounter the inverse of the structure predictions problem: How do I design sequences with a particular structure? Suppose, e. g., you have found a set of structural constraints necessary for a particular function, how do you know that these constraints are sufficient? The most stringent test would be to design and test sequences that are as random as possible, given the constraints.

Inverse folding can be viewed as an optimization problem that can be treated with simple heuristics. This is what the `RNAinverse` program does for you. Input for the `RNAinverse` program consists of an RNA secondary structure (the *target*) in bracket notation (on the first line), optionally followed by a sequence to be used as the starting point of the optimization (otherwise a random start sequence is used).

Suppose we need sequences that fold into the structure `(((((((((.....)))))...)))))` and we want the bulge sequence to be `CCA`.

1. Prepare an input file:

```
$ echo "(((((((((.....)))))...)))))\nNNNNNNNNNNNNNNNNNNNccaNNNN" > inverse.in
$ cat inverse.in
(((((((((.....)))))...)))))
NNNNNNNNNNNNNNNNNNNccaNNNN
$
```

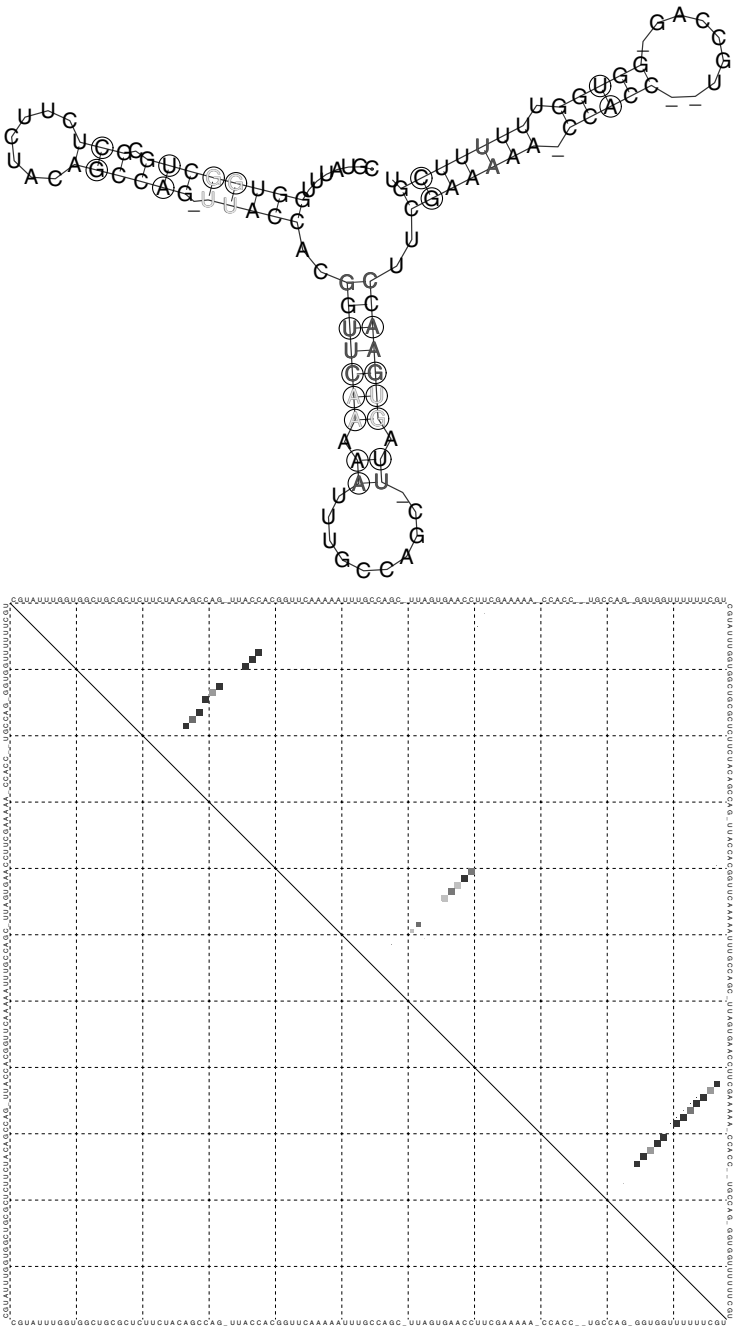


Fig. 12.2. Consensus structure and dot plot for the alignment of 9 RNAI sequences. Encircled bases in the consensus structure mark consistent and compensatory mutations.

The lower-case `cca` in the start sequence (second line of inverse.in) tells the program to keep these positions fixed; all other positions are random.

2. Design one sequence:

```
$  RNAinverse < inverse.in
GCGUACUAGGGAAUAGUccaAUGC    8
$
```

The 8 at the end of the line is the number of mutations done before a solution was found. Let us fold this sequence to check:

```
$  echo GCGUACUAGGGAAUAGUccaAUGC | RNAfold -p
GCGUACUAGGGAAUAGUCCAAUGC
(((((((((.....))))).....)))) ( -3.10)
((((,,,,.{{,.,,,)}}.)))) [ -3.91]
 frequency of mfe structure in ensemble 0.270197
$
```

The sequence indeed folds as desired. However, the output indicates that there are many alternative foldings, i. e., the structure is not well defined.

3. Design sequences with well defined structures, like this:

```
$ RNAinverse -Fmp < inverse.in
GCUCGGAUCGACUGUCCccaGGGU    7
GCCCCCUCUAAAUGAGGccaGGGC   12  (0.976354)
$ echo GCCCCCUCUAAAUGAGGccaGGGC | RNAfold -p
GCCCCCUCUAAAUGAGGCCAGGGC
(((((((((.....))))).....)))) (-11.00)
(((((((((.....))))).....)))) [-11.01]
 frequency of mfe structure in ensemble 0.976354
$
```

The options `-Fmp` tell `RNAinverse` to optimize the frequency of the MFE structure in the ensemble. The final sequence now has an extremely well-defined structure, as you can check by looking at the dot plot (`gv dot.ps`).
You can also design many sequences in one run by adding a `-R num` option to the command line.

12.3.4 Analysis of SELEX Experiments

Different aptamers from a SELEX experiment (see chapter 7) often evolve the same secondary structure to perform their function. However, it is often not possible to obtain good sequence alignment, and thus `RNAalifold` (or similar programs based on a sequence alignment) cannot be used to find the conserved structure.

In principle, the 'Sankoff algorithm' [7] can compute a consensus structure and an alignment simultaneously, but it is rarely used, due to its high computational cost. You can try `FOLDALIGN` [1], a simplified version of the algorithm that excludes multiloops. It is available as a web service at `http://www.bioinf.au.dk/FOLDALIGN/`.

Of course, you can also predict and compare individual structures.

1. Compute MFE structures:
   ```
   $ RNAfold < aptamers.seq > aptamers.fold
   ```
2. Compute all pairwise structure distances:
   ```
   $ RNAdistance -Xm < aptamers.fold > aptamers.dist
   ```
3. Do a cluster analysis to find sequences with similar structure:
   ```
   $ AnalyseDists -Xw < aptamers.dist
   $ gv wards.ps
   ```

Instead of `-Xw` you could choose `-Xn` to use 'nearest-neighbor joining' instead of Ward's clustering method (the postscript drawing will be called 'nj.ps'). In the resulting tree groups, sequences with similar structure can be easily recognized. You may then manually construct an alignment, to be processed by `RNAalifold`, analyze each cluster using `FOLDALIGN`, or try `pmmatch` (included on the CD).

Another alternative is to compute an alignment of the predicted MFE *structures*. The `RNAforester` program is a new tool to perform such structure alignments and can also construct multiple alignments (in contrast to `RNAdistance` above). It is available as a web service at `http://bibiserv.techfak.uni-bielefeld.de/rna-forester/` although it does not yet offer multiple alignments.

12.3.5 A Note for the Experts: Write your Own RNA Programs

The programs in the Vienna RNA package are all build on a code library that contains all the implemented algorithms. If you know some `C` programming you can use this library to develop your own programs. In addition, the package comes with a `Perl5` module that gives access to all the functions in the library from the scripting language `Perl`. As an example, the package contains a version of the RNAfold program written purely in Perl as well as the cgi script for an RNA folding web service.

12.4 Troubleshooting

The most frequent problems are related to input file formats. The sequence format used by the Vienna RNA package is very similar to FASTA, except that no line breaks or whitespace are allowed in the sequence. Line breaks will cause each

fragment to be folded separately (This is handy if you want to fold very many sequences, just put them all in one file, each on a separate line.)

Similarly, all programs read from `stdin` and write to `stdout`. Thus, the following does *not* work:

```
$ RNAfold input.seq
```

You *must* redirect input by writing:

```
$ RNAfold < input.seq
```

Often, the programs generate additional output files, such as a postscript drawing in the current directory. Therefore, change to a working directory where you have write permission before running any of the programs.

All Vienna RNA programs are documented in so-called man pages, a simple online help system for UNIX. To display a man page in your terminal, simply type `man` ***`program`***. Thus typing `man RNAfold` displays the full `RNAfold` manual with a detailed description of all program features. In contrast, typing `RNAfold -h` displays only a terse usage message. Be sure to read the man pages to obtain a full description of the available options.

12.5 Caveats

`Fundamental Rule:` Don't trust your computer any more than you trust your bench experiments.

Recall that the secondary-structure model for RNA is a model – and a crude one at that. It neglects pseudo knots and other tertiary interactions, does not take deviations from the additive 'nearest neighbor' energy model into account, and is based on thermodynamic parameters extracted from melting experiments by means of multidimensional fitting procedures. Thus, you cannot expect perfect predictions for each individual sequence. Rather, the accuracy is on the order of 50% of the base pairs for the minimum free energy structure.

In addition there is a wide variety of reasons why experimental facts and computer predictions disagree:

- The experimental work was performed at temperatures and/or ionic conditions far from those at which the RNA parameters were measured. Try the `-T` option to rescale the parameters for temperatures other than 37 °C.
- Your experiment was performed only with part of the sequence that you use on the computer or vice versa. RNA folding is an inherently nonlocal process. Whenever possible, use the exact same sequence on the computer and in the experiment.
- The molecule makes additional base pairs, either pseudoknots or bona fide 3D interactions.

- You compare the simulation to a published structure inferred from chemical probing. Often, probing data do not uniquely determine the structure. Unfortunately, there is no convenient tool (as yet) to compute all thermodynamically reasonable structures that are consistent with a collection of probing data.
- The RNA chain is caught in a metastable state in some molecules. This effect is crucial, e. g., for some RNA switches. Try `kinfold` if you suspect that the active RNA conformation might be a metastable state.
- Proteins that bind to the RNA may influence the folding. As a consequence, patterns of sequence co-variations that have evolved for RNAs that are functional in complexes with proteins (e. g., rRNA, RNAse P-RNA) might not conform very well with predicted folding for the isolated RNA.

When predicting consensus structures, be aware that the quality of the prediction is limited by the quality of the input alignment. Nucleic acid alignments are generally more error-prone than protein alignments. For coding sequences it is sometimes preferable to back-translate a protein alignment.

Acknowledgments

The sequences for the aptamer example included on the CD were kindly provided by U. Hahn. Financial support by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (P-15893) and the German DFG Bioinformatics Initiative is gratefully acknowledged.

References

1. J. Gorodkin, L.J. Heyer, and G.D. Stormo. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res.*, **25**:3724–3732, **1997**.
2. P.G. Higgs. RNA secondary structure: physical and computational aspects. *Q. Rev. Biophys*, **33**:199–253, **2000**.
3. I.L. Hofacker. The Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**:3429–3431, **2003**.
4. I. L. Hofacker, M. Fekete, and P.F. Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**:1059–1066, **2002**.
5. I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures (the Vienna RNA Package). *Monath. Chem.*, **125**:167–188, **1994**.
6. D. Mathews, J. Sabina, M. Zucker, and H. Turner. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, **288**:911–940, **1999**.

7. D. Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**:810–825, **1985**.
8. A. Schwienhorst. Structure-function analysis of RNAs generated by *In Vivo* and *In Vitro* selection. *Z. Phys. Chem.*, **216**:1–17, **2002**.
9. S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**:145–165, **1999**.
10. M. Zuker. Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, **10**:303–310, **2000**.

13 Patenting in Evolutionary Biotechnology

Martina Leimkühler and Hans-Wilhelm Meyers

13.1 Introduction

The biotechnology industry utilizes to a great extent the research results created by universities and other publicly funded research centers. Over the past years, these institutions have realized the value of their intellectual capital and have started to protect it by filing patent applications instead of merely publishing it as scientific articles and at conferences. The technology covered by such patent rights is then generally licensed to the industry for further development and commercialization. To protect the value of their inventions, researchers need to have at least a basic understanding of the major patenting aspects. In the following, some guiding principles for securing intellectual property in the biotechnological field are outlined.

13.2 The Nature of Patents

The fundamental concept underlying patent protection is simple: A patent allows its owner to prevent others from making commercial use of the patented invention without the owner's permission. The duration of this exclusive right is in general 20 years from the date of application. In the pharmaceutical field, an extension of up to 5 years is possible in many countries. For example in Europe, such an extension of the term is called a 'supplemental protection certificate' and is granted on request, if the patent protects a medicament that is on the market. A patent is limited to the territory of the state granting said patent.

It is a common misunderstanding that holding a patent also gives the right to practice the corresponding invention. However, this is not so: the rights given by a patent do not encompass the right to practice the invention, but only the right to keep others from doing so. Legislation might for instance interfere with the patentee's freedom to practice his invention. A typical example can be found in the field of pharmaceutical inventions: the patent owner for a new medicament needs permission from the public health authorities to market such medicament. The patentee's so called 'freedom-to-operate' might however also be restricted due

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst
Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-30799-0

to the existence of other patents. It is common for a first patentee to hold a patent for a basic invention that is the subject of further research and development activities resulting in patents for improvements to the basic invention. If the patent for such an improvement is held by someone other than the first patentee, then this second patent holder cannot practice the improvement without permission of the first patentee. Likewise, the first patentee cannot carry out the improvement without the consent of the other patent owner.

13.3 What Can Be Patented

In most countries, patents are granted for particular merits in the field of technology. Normally, the subject matter that can be protected by patents is limited to technical subjects such as devices, compositions of matter, or substances or to processes such as processes for manufacturing a product and work processes. If algorithms are the only subject matter for which protection is sought, there are often restrictions. In some countries, like the United States, however, patents are also granted with respect to subject matter of nontechnical background, for example, business methods or software. The challenging questions with respect to patenting of such business methods or patenting of software (including bioinformatics) is not dealt with in this chapter.

Numerous patent applications and patents exist which are directed to different aspects of evolutionary biotechnology. Commensurate with the contribution an invention makes to the state of the art in this technical field, different claim categories are represented in such intellectual property rights. A fundamental idea of the evolutionary design of molecules with predefined properties and functions consists in the application of consecutive selection cycles. Each such cycle typically comprises the three phases of (1) amplification, (2) diversification, and (3) selection. Claims may, for instance, be directed to the carrying out of such cycles under specific conditions. These may comprise defined mutation conditions as a source of variation; and an example of corresponding claims can be found in European patent EP 0 583 265 B1. The focus may however also be directed to the type of selection mechanism or to predefined selection criteria with regard to the fitness of the created molecule. A typical example is the SELEX (systematic evolution of ligands by exponential enrichment) technique, for which initial patent applications (for example, WO 91/19813) and a numerous number of follow-up patent applications directed to variations and improvements of the SELEX process have been filed. For example, U.S. patent 5,707,796 describes the use of the SELEX process in conjunction with gel electrophoresis to select nucleic acid molecules with specific structural characteristics. Other patents are directed to the selection of nucleic acids containing photoreactive groups (for example, U.S. patent 5,763,177) or to the Counter-SELEX method for identifying highly specific nucleic acid ligands able to discriminate between closely related molecules (for example, U.S.

patent 5,580,737), to mention just a few. A further typical claim category relates to products. Such products may comprise specific substances to be applied in a work process – such as a nucleic acid ligand with a photoreactive group which can be used in the above-described modified version of the SELEX process (see U.S. patent 5,763,177). Also substances identified from applying a specific work process may be patented – for example, ligands identified by the Counter-SELEX process described in U.S. patent 5,580,737. In summary, all typical claim categories can be found in intellectual property rights in evolutionary biotechnology.

The subject matter of an invention is disclosed in a section of the patent application that is called the description or specification. This section is followed by another section in which the claims are listed. The claims define for what matter protection is sought and define the scope of the exclusive right. The specification has to enable the skilled person to practice the invention.

The subject matter of a patent is the invention. It is generally accepted that an invention is a technical teaching for solving a real-world problem. Any invention starts with the knowledge and technology of the prior art, which normally poses a problem to be solved. The technical teaching, that is, the invention that is disclosed in the patent application, has to solve such problem. If the technical teaching of the invention is patentable, a patent will be granted. The most important essentials for assessing the patentability are novelty, inventiveness, industrial applicability, and sufficiency of disclosure of the invention. In many countries a patent application is granted only after an examination by patent offices. In some other countries only a very superficial examination is performed, if any. The validity of a granted patent can be challenged during or prior to litigation, typically by a party threatened by that patent.

13.4 The Requirement of Novelty

Novelty is essential to patentability and is well-defined in the respective patent laws of at least the United States of America, the European Patent Convention (EPC) [1], Japan, and other jurisdictions. Nothing can be patentable that is not new. Although novelty is basic to patentability, different concepts of novelty exist throughout the different patent systems in the world. The most straightforward is that of ‘absolute novelty’, which is applied by the European Patent Convention in Art. 54:

1. An invention shall be considered to be new if it does not form part of the state of the art.
2. The state of the art shall be held to comprise everything made available to the public by means of a written or oral description, by use, or in any other way, before the date of filing of the European patent application.

This means that an invention is considered new only if it does not form part of the broadly defined public state of the art before the date of filing of the European patent application (or the corresponding priority-establishing application). The state of the art in the sense of the EPC is not locally restricted. When examining the claims of the patent application with regard to their novelty, the European Patent Office considers prior publication of the invention irrespective of whether the publication occurred in one of the member states of the EPC or elsewhere.

In contrast to the EPC, the statutory standard for novelty in the United States, as set out in 35 U.S.C. § 102, gives a negative definition of novelty:

> A person shall be entitled to a patent unless (a) the invention was known or used by others in this country, or patented or described in a printed publication in this or a foreign country, before the invention thereof by the applicant for patent, or (b) the invention was patented or described in a printed publication in this or a foreign country or in public use or on sale in this country, more than one year prior to the date of the application for patent in the United States ...

The different definitions of novelty in these important patent systems have far-reaching consequences. An invention may be novel according to U.S. law, even if it lacks absolute novelty as applied by the EPC. For example, if an inventor describes the invention in a printed publication, he must apply for a patent in the United States before one year (according to § 102 (b)) has passed – otherwise any right to a U.S. patent is lost. In contrast, the inventor must file his patent application on the publication date, at the latest, if he wants to secure patent protection under the European Patent Convention and the national patent laws of many other countries.

What can we learn from these patent law stipulations for day-to-day business?

Regarding absolute novelty, it is recommended that inventors file a patent application, even if merely a provisional one, before samples constituting their invention – for example, an optimized enzyme, a vector, or other biological materials – are distributed, conference presentations are given, papers in a journal are published, or other disclosures are made. If, for whatever reason, filing of a patent application is delayed, inventors should distribute a sample only under a material transfer agreement protecting the confidential nature of the sample. Likewise, presentations of the invention should be given only under a confidentiality agreement. Failure to comply with the absolute novelty requirement may disastrously affect obtaining a valid patent protection in essentially all industrial countries other than the United States, Canada, Australia, and Japan. These states are among the few countries that still have a grace period. In contrast to the 12-month period granted by the U.S., Canadian, and Australian law, the Japanese grace period has a term of only 6 months.

How is novelty assessed by the patent authorities?

As explained above, most patent authorities consider an invention as novel when its subject matter, as defined in the claims, was not disclosed prior to the first filing date of the invention. The claims are a listing of features defining the invention in certain categories and usually in generalized terms which have to be supported by the description or specification. The most important patent categories are devices, substances, processes, and method of use. For example, if a publicly available reference discloses each and every feature of an invention as claimed, then the invention is no longer novel. The reference can be a publication in oral, written, electronic, or any other form. However, if more than one reference is necessary to identify each and every feature of a claim, then the subject matter of the claim is novel. During the patenting process one can amend the claims to avoid anticipation by references. Any claim amendment, however, has to be supported by the description or the specification or has to be disclosed literally. Any feature that is added to a claim or any amendment of a feature in a claim has to be disclosed in the specification, at least precise enough that no 'new matter' is introduced into the patent application.

What is a priority right?

Any patent application that was filed in accordance with the patent law of the country in which it was filed is the basis for a so-called 'priority right'. Most countries of the World Trade Organization (WTO) are also members of the Paris Convention. The countries of the Paris Convention grant each other the right to claim the application date of an application that was filed earlier in one of the member states. The later filing, however, has to be done within one year of the first filing. This has the effect that any prior art that was published between the first filing of the invention in one of the Paris Convention states and the actual filing in the respective other country is not regarded when assessing the patentability, unless the application filed later is not identical with the one filed first. Priority right is a very important tool in patent prosecution.

If the second (later) application claims subject matter that was not disclosed in the first application, any prior art published before the filing date of the later application is considered. Thus, such newly introduced subject matter has to be both novel and inventive with respect to the prior art published before the later filing date. Normally, this does not cause problems. However, if a researcher had filed a first patent application before the respective scientific manuscript on his invention was published and thereafter files (within the priority year) a second application claiming an improved embodiment of the invention, then the improvement not only has to be novel, but also inventive with respect to the subject matter of the manuscript published between the two filing dates. Very often, the improved embodiment is not inventive over the published manuscript. Thus, the improvement is not patentable over the disclosure of the invention in the published manuscript, although the content of the publication was filed as the first patent application. Therefore, publishing the invention within the priority year should be avoided.

13.5 The Requirement of Inventiveness

Although novelty is a well-defined issue, the question of inventiveness (the presence of an inventive step in EPC terminology) is often more striking. When novelty has been acknowledged by an authorized examining authority, a patent is granted only if the subject matter is not obvious, considering the prior art. Article 56 EPC defining the inventive step reads as follows:

> An invention shall be considered as involving an inventive step if, having regard to the state of the art, it is not obvious to a person skilled in the art ...

The respective U.S. regulation is set out in 35 U.S.C. § 103(a), as follows:

> A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

The references cited by an examining authority have to be assessed in view of the so-called 'person skilled in the art'. The skilled person is a fiction and represents a person who knows all references that have ever been published in whatever language, who however does not have too much creativity in posing and solving objects or problems on the respective technical field. This means that the person skilled in the art does not have the capacity of a Nobel Laureate, but does know more about the respective technical field than a technically interested layman. Of course, the knowledge of the person skilled in the art is stretchable and a matter of the respective case. Occasionally, a team of persons can be addressed as 'person skilled in the art', in particular in border-line technical fields.

The issue of inventiveness is more complicated and ambiguous than the issue of novelty. It depends for example on the state of the art, the skill of the person skilled in the art, whether the invention uses well-known techniques, whether the invention solves the problem in the prior art only in a further way employing techniques that are known as such, and other issues.

For example, in the early 1980s it was common to obtain a patent on a full-length protein although its sequence was partly known, for instance by disclosure of the protein's N-terminal amino acid sequence. However today, with our knowledge of the arsenal of methods and techniques in biochemistry and molecular biology, in most instances no inventiveness is involved in elucidating such a full-length sequence. The skilled person would readily be able to translate the partial amino acid sequence into the respective nucleic acid sequence to use this as a probe. Knowing about the genetic code, it is a routine procedure to identify a cDNA and consequently the full-length amino acid sequence of the protein.

Inventiveness has to be shown by the applicant or the patentee, if a lack of inventiveness is claimed by the patent examiner during the prosecution or the public in an opposition procedure. An indication of inventiveness is, for example, an unexpected result proven by comparative examples or improved properties of a new substance. Thus, if it can be shown that a new peptide or protein shows improved properties over the wild-type protein, inventiveness can usually be acknowledged. When aiming to achieve patent protection for evolutionarily designed molecules with predefined properties and functions, an inventor should keep this in mind.

13.6 The Requirement of Utility

Patent systems typically require that the claimed invention must have utility (as set forth in 35 U.S.C. § 101) or must be susceptible to industrial application (as set forth in Art. 57 EPC). With regard to genetic patenting, the European Patent Convention explicitly states that the industrial application of a sequence or a partial sequence of a gene must be disclosed in the patent application (see Rule 23e(3) EPC)). The requirement to associate the sequence with a function can also be found in U.S. law.

13.7 The Requirements of Enablement and Written Description

Apart from the essentials of novelty, unobviousness, and utility discussed above, further challenges require special attention when seeking patent protection. The invention should be described in such a manner as to comply with both the written-description and the enablement requirement. These requirements are contained in 35 U.S.C. § 112, which states that "the specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same . . . ". Similar regulations exist in other patent laws, as exemplified by Articles 83 and 84 EPC.

An underlying concept of the patent system is to grant a patent owner an exclusive right for the commercialization of his invention subject to meeting certain requirements, one such requirement being that the patent application sufficiently discloses the invention to the public. It is ensured by requiring sufficient enablement that the person skilled in the art can practice the full scope of the claimed invention. The invention should be disclosed in such a way that the skilled person does not need undue experimentation for practicing it. Claims are usually drafted in broad terms to keep others from circumventing or designing around a patented

invention. Such broad claims should be sufficiently supported by the patent specification. This means that the breadth of the claims and the scope of enablement given by the patent specification should correspond to each other. Otherwise, hurdles in the patent prosecution process will be inevitable, and even if a patent is issued at last, it might be ruled invalid when trying to enforce it.

To satisfy the written-description requirement, the invention must be described in sufficient detail. A person of ordinary skill in the art should recognize that the inventor was in possession of the claimed invention at the time of filing. The requirements of written-description and enablement are distinct from each other. This means that, although a patent specification may sufficiently enable a person skilled in the art to practice the invention, it may still be found to not meet the written-description requirement. The case "*The Regents of the University of California v. Eli Lilly and Co.*" [2] provides an example for the significance of the written-disclosure requirement. The disclosure contained the nucleotide sequence of a rat proinsulin cDNA and a general method for obtaining the corresponding human cDNA. The patent claims, however, were broadly drafted and covered not only the rat cDNA but also vertebrate, mammalian, and human cDNA. These were held invalid because the specification did not provide an adequate written description.

A substantial amount of intellectual property generated in the biotechnological field relates to research tools. Examples of research tools are numerous and include, for instance, evolutionary optimization processes for generating biopolymers with improved properties and assays to screen compound libraries for potential drug candidates as well as for novel disease targets such as receptors or ion channels. These valuable research tools often smooth the way for making downstream products of even greater commercial value. In some instances, inventors have tried to participate in the value of such downstream products by drafting so-called 'reach-through' claims.

The scope of a reach-through claim goes beyond that of the actual research tool and is directed to the downstream product or its uses. A typical example is the identification and characterization of a receptor molecule that can be used in a screening assay to identify novel drug candidates. A reach-through claim would be directed to compounds identifiable from the carrying out of such a screening assay without even actually having identified or disclosed any specific compounds in the patent specification. In such a case, the claims covering the compounds solely rely on the identification of the disease target and its use in the corresponding screening assay. Further examples may also be found in the field of evolutionary biotechnology, for instance, if reach-through claims are directed to optimized molecules putatively identifiable from the carrying out of an evolutionary improvement process.

In the recent case "*University of Rochester v. G.D. Searle & Co.*", claims of the reach-through type were at issue [3]. Scientists of the University of Rochester cloned the gene that produces PGHS-2 (also known as Cox-2) in the early 1990s. This enzyme promotes pain and inflammation. In contrast, the known enzyme PGHS-1 (or Cox-1) is beneficial, as it helps protect the stomach lining. Claims were directed to a method for selectively inhibiting PGHS-2 activity in humans by administering a nonsteroidal compound. When the patent was issued in 2000, several drug

companies had already developed and successfully marketed Cox-2 inhibitors. The companies were sued by the University of Rochester, seeking damages for alleged patent infringement. In the university's opinion, the basic findings of its researchers had paved the way for these companies.

In the trial, the court had a close look at the patent specification, which disclosed an assay for identifying compounds that inhibit PGHS-2. Compounds that may possibly be found by conducting the screening assay could be used in the treatment of pain. It was postulated that these should not have the typical undesirable side effects, such as stomach irritation associated with widely used pain relievers. In addition, the specification identified some broad categories of compounds that might work as such drugs. In the court's opinion, the claimed method of treatment "depends upon finding a compound that selectively inhibits PGHS-2 activity. Without such a compound, it is impossible to practice the claimed method of treatment. It means little to 'invent' a method if one does not have possession of a substance that is essential to practicing that method." Consequently, the U.S. district court ruled that the patent was invalid, for failure to meet both the written-description and the enablement requirement. The decision has been appealed by the University of Rochester.

In summary, the general tenor in U.S. decisions [2–4] is that it is not sufficient for obtaining sound patent protection to merely disclose a wish or a plan for obtaining the claimed invention. Reach-through claims are subject to thorough scrutiny. An inventor should provide evidence that he had, at the time of filing a patent application, a complete conception of the downstream product that is being claimed. An inventor might, for example, describe a specific structural element of the downstream product in the specification to give such evidence. The enablement and written-disclosure requirement should be carefully addressed by inventors in the biotechnology field by providing a well-founded specification.

13.8 Patent Prosecution

When seeking patent protection, an inventor typically files a national patent application in his country of residence. Once this first patent application has been filed, the priority year provided for by the Paris Convention starts to run. During the priority year, an inventor will normally continue his work on the invention, for instance, by conducting further experiments. All this material can be used in preparing patent applications to be filed abroad.

One important tool in the international patent world is the Patent Cooperation Treaty (PCT). By filing an 'international' patent application, one may seek patent protection for an invention simultaneously in each of a large number of countries. Such an application may be filed by anyone who is a national or resident of a contracting state of the PCT. A list of the more than 120 contracting states can be found on the website of the World Intellectual Property Organization (WIPO) [5]. The effect of the international application in each such state is the same as if a

national patent application had been filed with the national patent office of that state. Specific regulations may apply if the designated state is also a party to a regional patent convention such as the European Patent Convention (EPC). It is important to understand that the prosecution process under the PCT does not comprise the grant of a patent. This is the responsibility of the national or regional patent authorities, as explained in more detail below.

The international patent application is subjected to an 'international search'. This search is carried out by one of the major patent offices, such as the U.S. Patent and Trademark Office, the Japanese Patent Office, or the European Patent Office. The search results in an 'international search report' (ISR). An example of such a report is shown in Figure 13.1. The international search report contains a classification of the subject matter of the invention according to an international patent classification system (IPC) and an identification of the fields searched. The heart of the ISR, however, is a list of the citations of such published documents that might affect the patentability of the invention claimed in the international patent application. The cited documents are categorized according to their nature. For example, in view of a document of category *X*, the claimed invention cannot be considered novel or cannot be considered to involve an inventive step. Whereas the patentability of the claimed invention is severely questioned on basis of an *X* document taken alone, in the *Y* category the concepts of several documents have to be combined. The claimed invention cannot be considered to involve an inventive step when a document of category *Y* is combined with one or more such documents. Such combination however must be obvious to a person skilled in the art. Further categories relate to documents defining the general state of the art (*A* category) or documents published within the priority year, that is, prior to the international filing date but later than the priority date claimed (*P* category). As discussed above, a *P* document may become highly relevant if the claimed subject matter of the international patent application goes beyond the scope of the priority application. The claims might not be entitled to the original priority any more. Therefore, an inventor should not publish his invention before the end of the priority year, if possible under the specific circumstances.

The international search report is communicated to the applicant by the international search authority. The applicant may then decide to withdraw his patent application if, for example, the prior art found makes the granting of a patent highly unlikely. If the international patent application is not withdrawn, it is published together with the search report. Third parties now have the possibility to take notice of the claimed invention and to form their own opinion about its presumed patentability. In addition, they can assess any possible later dependencies should they want to practice the invention. An example of the cover of a published international patent application is shown in Figure 13.2.

The applicant may additionally ask for an 'international preliminary examination report' (IPER). This report gives a preliminary, nonbinding opinion on the patentability of the claimed invention. On the basis of the international preliminary examination report, the applicant can once again evaluate the chances of his invention being patented. The applicant is entitled to amend the international application during the international preliminary examination. He might for example

INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP92/00840

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl^5 : C12Q 1/68; C12N 15/10; C12P 21/00
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED
Minimum documentation searched (classification system followed by classification symbols)
Int.Cl^5 : C12Q; C12N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	SCIENCE volume 249, 3 August 1990, LANCASTER, PA US pages 505 - 510; C. TUERK ET AL.: "Systematic evolution of ligands by exponential enrichment : RNA ligands to bacteriophage T4 DNA polymerase" (cited in the application) see the whole document ---	1
A	BERICHTE DER BUNSEN-GESELLSCHAFT FÜR PHYSIKALISCHE CHEMIE volume 89, 1985, WEINHEIM, DE pages 658 - 667; M. EIGEN : "Macromolecular evolution : dynamical ordering in sequence space" see the whole document, in particular abstract and "Conclusions" ---	1
A	NATURE. volume 344, 29 March 1990, LONDON GB pages 467 - 468; ./.	1

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13 August 1992 (13.08.92)	16 September 1992 (16.09.92)
Name and mailing address of the ISA/ European Patent Office Facsimile No.	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (July 1992)

Fig. 13.1. Example international search report listing several prior-art documents. These are categorized with respect to their influence on the patentability of the subject matter of the claims.

INTERNATIONAL SEARCH REPORT

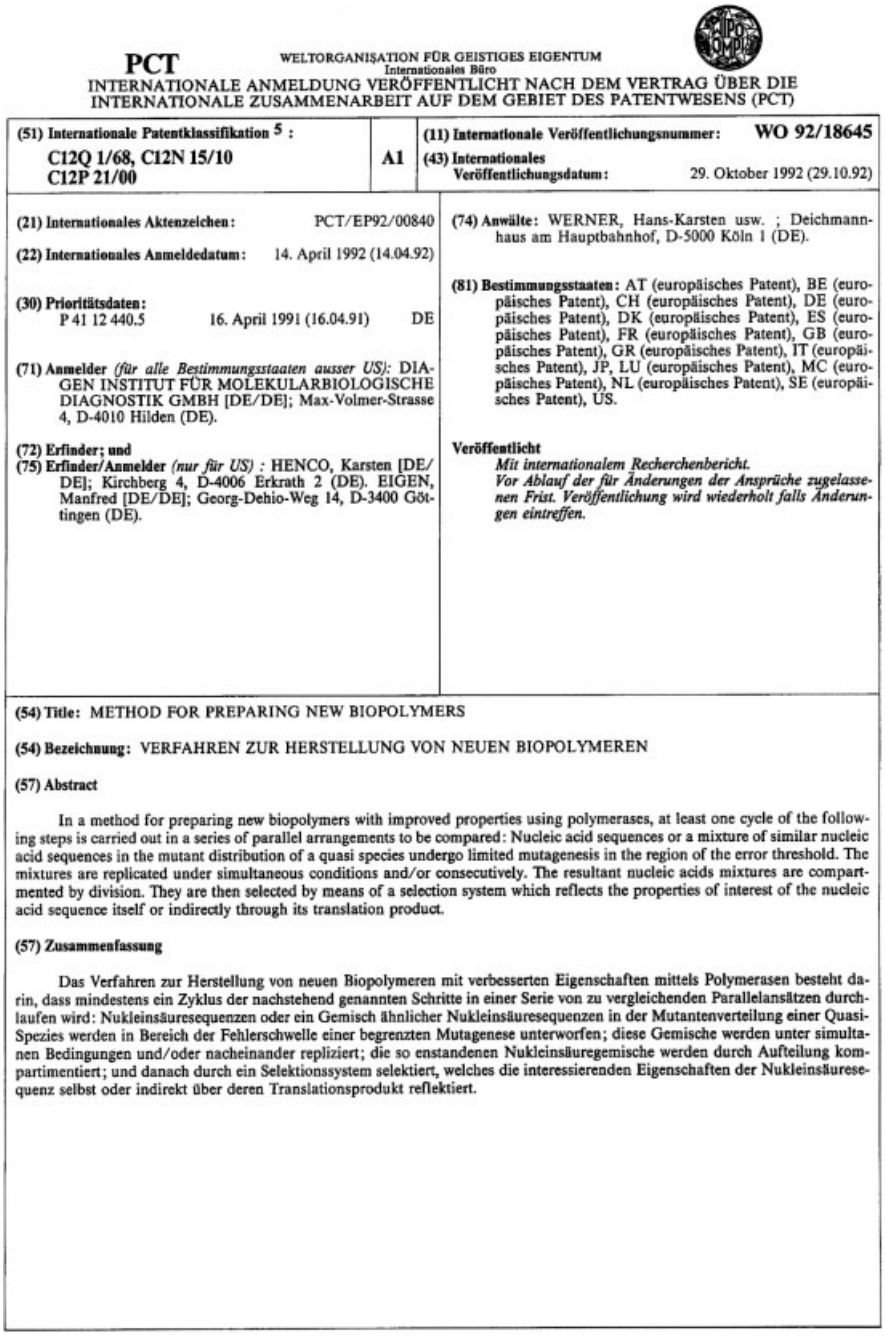
International application No.
PCT/EP92/00840

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	D.L.ROBERTSON ET AL.: "Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA" see the whole document	
A	GENE. volume 82, 1989, AMSTERDAM NL pages 83 - 87; G.F. JOYCE : "Amplification, mutation and selection of catalytic RNA" see the whole document	1
A	EP, A, 0285123 (SUOMEN SOKERI OY) 5 October 1988 see page 7, line 50 - page 8, line 14; claims	1
P,X	WO, A, 9105058 (KAWASAKI G.) 18 April 1991 see page 2, line 37 - page 7, line 21 see page 24, line 36 - page 25, line 11; claims	1,8
E	WO, A, 9202536 (THE REGENTS OF THE UNIVERSITY OF COLORADO) 20 February 1992 see the whole document in particular page 2, line 30 - page 15, line 11, page 40, line 6 - page 41, line 2 and claims	1,8

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

Fig. 13.1 (Continued).



PCT WELTORGANISATION FÜR GEISTIGES EIGENTUM
Internationales Büro
INTERNATIONALE ANMELDUNG VERÖFFENTLICHT NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT)

(51) Internationale Patentklassifikation 5 : C12Q 1/68, C12N 15/10 C12P 21/00 — A1

(11) Internationale Veröffentlichungsnummer: WO 92/18645

(43) Internationales Veröffentlichungsdatum: 29. Oktober 1992 (29.10.92)

(21) Internationales Aktenzeichen: PCT/EP92/00840

(22) Internationales Anmeldedatum: 14. April 1992 (14.04.92)

(30) Prioritätsdaten:
P 41 12 440.5 16. April 1991 (16.04.91) DE

(71) Anmelder *(für alle Bestimmungsstaaten ausser US)*: DIAGEN INSTITUT FÜR MOLEKULARBIOLOGISCHE DIAGNOSTIK GMBH [DE/DE]; Max-Volmer-Strasse 4, D-4010 Hilden (DE).

(72) Erfinder; und
(75) Erfinder/Anmelder *(nur für US)* : HENCO, Karsten [DE/DE]; Kirchberg 4, D-4006 Erkrath 2 (DE). EIGEN, Manfred [DE/DE]; Georg-Dehio-Weg 14, D-3400 Göttingen (DE).

(74) Anwälte: WERNER, Hans-Karsten usw. ; Deichmannhaus am Hauptbahnhof, D-5000 Köln 1 (DE).

(81) Bestimmungsstaaten: AT (europäisches Patent), BE (europäisches Patent), CH (europäisches Patent), DE (europäisches Patent), DK (europäisches Patent), ES (europäisches Patent), FR (europäisches Patent), GB (europäisches Patent), GR (europäisches Patent), IT (europäisches Patent), JP, LU (europäisches Patent), MC (europäisches Patent), NL (europäisches Patent), SE (europäisches Patent), US.

Veröffentlicht
Mit internationalem Recherchenbericht.
Vor Ablauf der für Änderungen der Ansprüche zugelassenen Frist. Veröffentlichung wird wiederholt falls Änderungen eintreffen.

(54) Title: METHOD FOR PREPARING NEW BIOPOLYMERS

(54) Bezeichnung: VERFAHREN ZUR HERSTELLUNG VON NEUEN BIOPOLYMEREN

(57) Abstract

In a method for preparing new biopolymers with improved properties using polymerases, at least one cycle of the following steps is carried out in a series of parallel arrangements to be compared: Nucleic acid sequences or a mixture of similar nucleic acid sequences in the mutant distribution of a quasi species undergo limited mutagenesis in the region of the error threshold. The mixtures are replicated under simultaneous conditions and/or consecutively. The resultant nucleic acids mixtures are compartmented by division. They are then selected by means of a selection system which reflects the properties of interest of the nucleic acid sequence itself or indirectly through its translation product.

(57) Zusammenfassung

Das Verfahren zur Herstellung von neuen Biopolymeren mit verbesserten Eigenschaften mittels Polymerasen besteht darin, dass mindestens ein Zyklus der nachstehend genannten Schritte in einer Serie von zu vergleichenden Parallelansätzen durchlaufen wird: Nukleinsäuresequenzen oder ein Gemisch ähnlicher Nukleinsäuresequenzen in der Mutantenverteilung einer Quasi-Spezies werden in Bereich der Fehlerschwelle einer begrenzten Mutagenese unterworfen; diese Gemische werden unter simultanen Bedingungen und/oder nacheinander repliziert; die so enstandenen Nukleinsäuregemische werden durch Aufteilung kompartimentiert; und danach durch ein Selektionssystem selektiert, welches die interessierenden Eigenschaften der Nukleinsäuresequenz selbst oder indirekt über deren Translationsprodukt reflektiert.

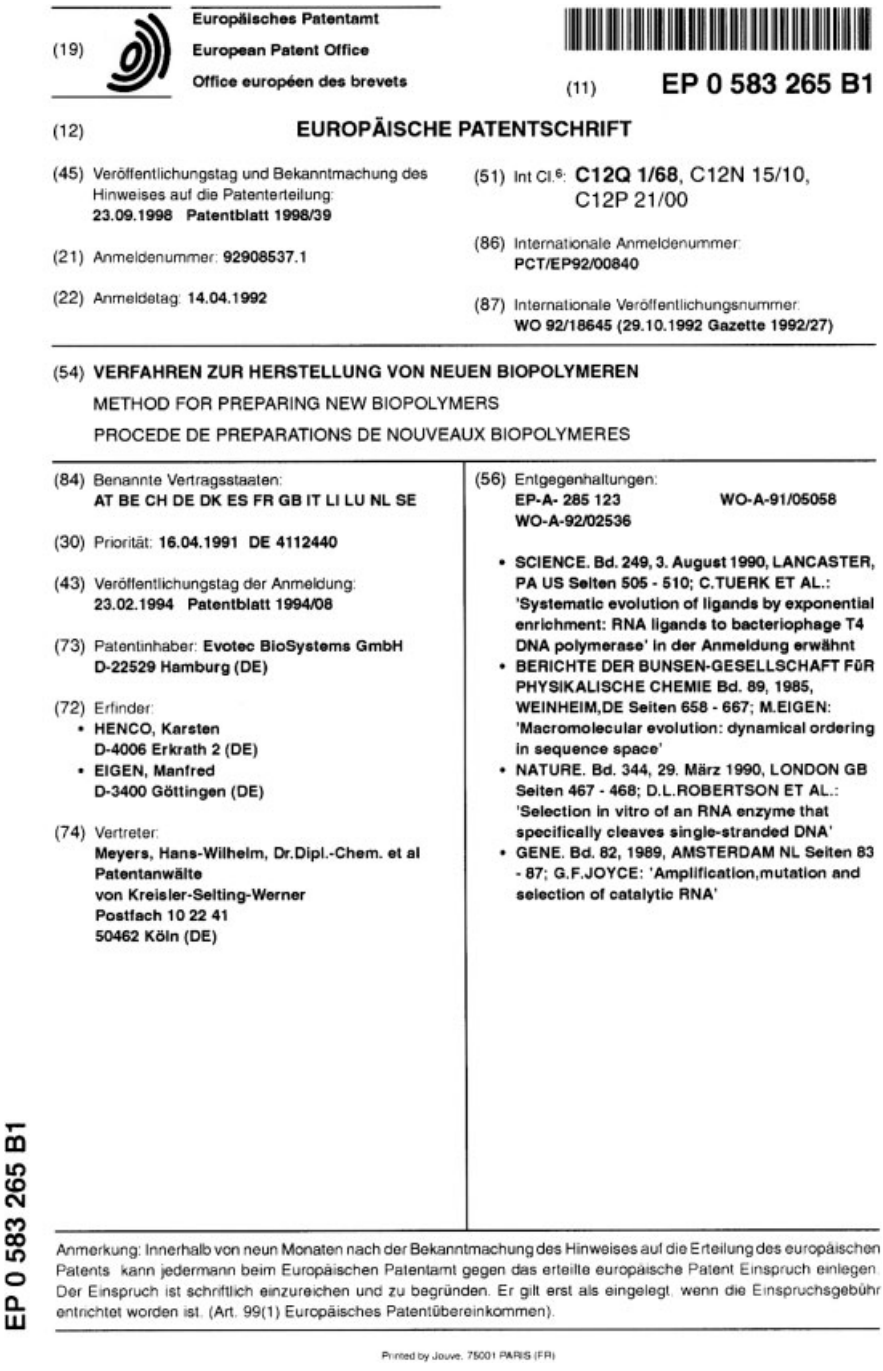
Fig. 13.2. Example cover for an international patent application as published. It is denoted an A-document.

restrict the scope of the patent claims due to the prior art on record. Of course, he may not add 'new matter'. Modifications to the PCT system have been introduced from 1 January, 2004. Each international search authority now prepares an International Search Report already supplemented with a preliminary written opinion on patentability (the so-called PCT International Search Opinion). Upon applicant's request, this opinion is subject to a dialogue with the patent examiner of the international preliminary examination authority. As mentioned above, the prosecution process under the PCT does not constitute the grant of a patent. This is the responsibility of the national or regional patent authorities. If the applicant decides to continue with the international application, he must commence a so-called national (or regional) phase before the corresponding national or regional patent offices. If necessary, the applicant must furnish a translation of the application into the official language of that office. In addition, he must pay the usual fees to the office. Most such patent offices then reexamine the patent application with regard to the specific patentability requirements discussed above. If these are met, a patent is granted. An example of the cover of a granted European patent is shown in Figure 13.3. Although the publication document of a European patent application (as well as the publication under the PCT) is designated with the letter A, a granted European patent can be identified by the letter B. Formerly, the U.S. published only granted patents (without using any specific letter code). Today, publication of U.S. patent applications is required by the American Inventors Protection Act of 1999. Patent applications filed in the U.S. on or after November 29, 2000, are published as 'A-documents' and can be distinguished from the corresponding granted patents, published as 'B-documents'. As a result of publication of a patent application, an applicant may assert provisional rights. The patentee may obtain a reasonable financial consideration from a third party that makes commercial use of the claimed subject matter of a published patent application. For this purpose, the patentee must have met certain requirements and, of course, the patent issued from the application must have a substantially identical claim.

Looking at the hurdles in the patenting processes, one may wonder how inventors can improve their chances of obtaining valid patent rights. Below, some suggestions are made.

13.9 Search Tools

As discussed in detail above, essential patentability requirements are novelty and inventiveness with respect to the prior art. In filing a patent application, it is therefore desirable that its content as claimed is at least novel with respect to the prior art. The question of whether an invention is obvious, considering the prior art, is often difficult to assess and should be subject to discussion with the corresponding patent authority. To avoid 'inventing' subject matter already included in the prior art, inventors should have a clear knowledge and understanding of the state of



(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 583 265 B1

(12) **EUROPÄISCHE PATENTSCHRIFT**

(45) Veröffentlichungstag und Bekanntmachung des Hinweises auf die Patenterteilung:
23.09.1998 Patentblatt 1998/39

(51) Int Cl.[6]: **C12Q 1/68**, C12N 15/10, C12P 21/00

(21) Anmeldenummer: **92908537.1**

(86) Internationale Anmeldenummer:
PCT/EP92/00840

(22) Anmeldetag: **14.04.1992**

(87) Internationale Veröffentlichungsnummer:
WO 92/18645 (29.10.1992 Gazette 1992/27)

(54) **VERFAHREN ZUR HERSTELLUNG VON NEUEN BIOPOLYMEREN**
METHOD FOR PREPARING NEW BIOPOLYMERS
PROCEDE DE PREPARATIONS DE NOUVEAUX BIOPOLYMERES

(84) Benannte Vertragsstaaten:
AT BE CH DE DK ES FR GB IT LI LU NL SE

(30) Priorität: **16.04.1991 DE 4112440**

(43) Veröffentlichungstag der Anmeldung:
23.02.1994 Patentblatt 1994/08

(73) Patentinhaber: **Evotec BioSystems GmbH**
D-22529 Hamburg (DE)

(72) Erfinder:
- **HENCO, Karsten**
  D-4006 Erkrath 2 (DE)
- **EIGEN, Manfred**
  D-3400 Göttingen (DE)

(74) Vertreter:
Meyers, Hans-Wilhelm, Dr.Dipl.-Chem. et al
Patentanwälte
von Kreisler-Selting-Werner
Postfach 10 22 41
50462 Köln (DE)

(56) Entgegenhaltungen:
EP-A- 285 123 **WO-A-91/05058**
WO-A-92/02536

- **SCIENCE. Bd. 249, 3. August 1990, LANCASTER, PA US Seiten 505 - 510; C.TUERK ET AL.: 'Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase' in der Anmeldung erwähnt**
- **BERICHTE DER BUNSEN-GESELLSCHAFT FüR PHYSIKALISCHE CHEMIE Bd. 89, 1985, WEINHEIM,DE Seiten 658 - 667; M.EIGEN: 'Macromolecular evolution: dynamical ordering in sequence space'**
- **NATURE. Bd. 344, 29. März 1990, LONDON GB Seiten 467 - 468; D.L.ROBERTSON ET AL.: 'Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA'**
- **GENE. Bd. 82, 1989, AMSTERDAM NL Seiten 83 - 87; G.F.JOYCE: 'Amplification,mutation and selection of catalytic RNA'**

Anmerkung: Innerhalb von neun Monaten nach der Bekanntmachung des Hinweises auf die Erteilung des europäischen Patents kann jedermann beim Europäischen Patentamt gegen das erteilte europäische Patent Einspruch einlegen. Der Einspruch ist schriftlich einzureichen und zu begründen. Er gilt erst als eingelegt, wenn die Einspruchsgebühr entrichtet worden ist. (Art. 99(1) Europäisches Patentübereinkommen).

Printed by Jouve, 75001 PARIS (FR)

EP 0 583 265 B1

Fig. 13.3. Example cover of a European patent as published. It is denoted a B-document.

the art. Scientific and patent databases provide an excellent tool for assessing the technical knowledge in the field of invention. For searching prior art in the field of evolutionary biotechnology, the following databases are very useful:

- PubMed (http://www.ncbi.nlm.nih.gov/), developed by the National Center for Biotechnology Information (NCBI), which includes (among other things):
  - Access to the bibliographic database Medline, which covers such fields as medicine, veterinary medicine, the health care system, and the preclinical sciences. Bibliographic citations and author abstracts from more than 4600 biomedical journals worldwide are included in this database. Most records are from English-language sources or have English abstracts.
  - Access to a biosequence database.
- Other sequence databases can be found via the website (http://www.ebi.ac.uk/) of the European Bioinformatics Institute (EBI). Databanks managed by the EBI include:
  - EMBL Nucleotide Database (collection of nucleotide sequences).
  - Swiss-Prot (annotated protein sequences).
  - ArrayExpress (gene expression data).
- The Publication Site for Issued and Published Sequences (http://seqdata.uspto.gov/).

The utilization of patent information is, of course, promoted by diverse patent authorities. The European Patent Office (http://www.european-patent-office.org/) provides a service called esp@cenet, which is easily accessible via the Internet (http://ep.espacenet.com/). In addition, inventors might conduct online file inspections via a service called epoline, which can be reached at http://www.epoline.org/. The user gains direct access to all published European patent applications and patents stored in electronic form. Patent information services are also offered by the United States Patent and Trademark Office (http://www.uspto.gov/).

13.10 The First-to-invent Principle of the United States and Its Consequences on Laboratory Notebook Keeping

From time to time, different persons invent the same subject matter and intend to protect it by patent rights. In these cases, two or more patent applications are filed by the inventors claiming substantially the same invention. Most countries in the world apply the first-to-file principle to determine who is entitled to the patent. In contrast, the United States do not give priority in case of conflict to the first applicant, but to the first to invent. Interference proceedings are instituted to determine who is the first inventor and consequently entitled to the patent (see 35 U.S.C. § 102 g (1)). The parties involved in such a proceeding may provide evidence of facts to prove their date of invention.

A second circumstance in which it may be necessary to prove a date of invention is set forth in 35 U.S.C. § 102 g (2) which states that "a person shall be entitled to

a patent unless before such person's invention thereof, the invention was made in this country by another inventor who had not abandoned, suppressed, or concealed it." In determining priority of invention, there shall be considered not only the respective dates of conception and reduction to practice of the invention, but also the reasonable diligence between conception and reduction to practice.

The complete mental realization of an invention is called conception. It is the formation, in the mind of the inventor, of "a definite and permanent idea of the complete and operative invention, as it is thereafter to be applied in practice" [4]. The idea must be "so clearly defined in the inventor's mind that only ordinary skill would be necessary to reduce the invention to practice, without extensive research or experimentation" [4].

The United States Patent and Trademark Office (USPTO) provides a service under the Disclosure Document Program to give evidence as to the date of conception of an invention [5]. An inventor may disclose the conception of his invention in a Disclosure Document to the USPTO for a nominal fee. This document will be held in confidence by the USPTO. After a period of two years it will be destroyed unless a related patent application is filed within these two years. They caution that "the Disclosure Document is not a patent application. The date of its receipt in the USPTO will not become the effective filing date of any patent application subsequently filed" [6].

After conceiving the invention, the inventor's next step is to reduce the invention to practice. A reduction to practice can be a constructive reduction to practice, which occurs when a patent application is filed. It can, however, also be an actual reduction to practice, which is the physical realization of the invention. For instance, in the case of a composition it includes the actual making thereof. With regard to a process, it includes the actual carrying out of the steps of the process. The determination that the invention will work for its intended purpose is usually also necessary.

In both academia and industry, the laboratory notebook is a legal document that records the work done by an individual researcher. Even in the USPTO's own view, its Disclosure Document Program "does not diminish the value of the conventional, witnessed, permanently bound, and page-numbered laboratory notebook or notarized records" [6]. Without a doubt, any experimental work might be worthless unless it is properly recorded for later use. Evidence of the dates of conception and reduction to practice of an invention may be established by a well-maintained laboratory notebook. Therefore, researchers should become familiar with how to keep persuasive laboratory notebooks. Of course, no notebook is ever perfect. However, the closer one comes to this objective, the better it will be.

The notebook is an evidentiary document. Consequently, it should be maintained in such a manner that it cannot be manipulated. If, for instance, pages can be exchanged or added, it would be very difficult to prove that a particular page was not inserted at some later date. Therefore, the notebook should have permanently bound pages that are consecutively numbered.

Researchers are often uncertain about the scope of description of their experiments to be included when preparing a lab notebook. First of all, it is important to realize that the laboratory notebook serves not only for corroboration of the reduc-

tion to practice of an invention but also of its conception. Therefore, one should include the goal or the idea behind an experiment. It should be clear why the data was generated. A skilled worker should be able to repeat the original work on the basis of the notebook entries. It is therefore important for all inventions, but perhaps especially for chemical and biotechnological ones, to describe in detail all experimental procedures. These should include in-depth instructions for performing the experiment. Information necessary to unambiguously identify the reagents (including source, purity, concentration, etc.) should be given, as well as the equipment used for conducting the experiment. All numbers must have units. Abbreviations or code names should not be used unless they are clearly defined in the notebook. Preferably, records such as photographs or printouts of analytical data should be permanently attached in the notebook. And last, the results of the experiment should be carefully recorded. One should always remember that a notebook must be factual. Consequently, statements such as 'the idea is obvious' or any other subjective language should be avoided.

Because the notebook serves as a document that provides evidence of dates of invention, entries should be made as soon as possible after conceiving an invention and reducing it to practice, for example, after a particular experiment has been performed. The entry has to be dated with the date it is made and be signed by the person making the entry. Entries should always be made in the notebook without skipping pages or leaving empty spaces. A line should be drawn through any unused portion of a page. It is extremely important that the notebook entries be witnessed [7, 8]. Someone who is not an inventor should read and understand the entries. The process of witnessing is typically done by a signature and date under a statement saying "read and understood" on each page of the notebook. Without a witness to sign the notebook, the notebook would be based solely on the testimony of the inventor. Between the date an experiment was performed and entered into the notebook and the date such entry was witnessed, there should not be a long time span. Witnessing should rather take place contemporaneously with the entry of the inventor. If changes to an entry must be made at a later date, any entry made in error should be crossed out. Erasures should be avoided. The changes should be signed and dated by the inventor and rewitnessed by the witness next to the correction.

Records are often kept in electronic format in many laboratories. Most computer records can be willfully updated and changed. Consequently, evidence that their content was created at a particular time can be severely questioned. Therefore, hard copies should be made of such electronic records. Such hardcopies should be signed, dated, and witnessed as described in detail above.

Last, the use of a laboratory notebook should be controlled, for example, by a central department. Under no circumstances should it be treated as a freely available publication. This would be in conflict with the above-discussed patentability requirements of novelty and inventiveness.

13.11 Summary

The purpose of this chapter is to highlight the importance and practical procedures of patenting in biotechnology. Special attention is drawn to the patentability requirements under the European and U.S. patent law. Practical guidelines are given to researchers aiming to protect their inventions by patents. This chapter should not, however, be used as a substitute for the advice of a lawyer or patent attorney taking into consideration the reader's specific circumstances.

References

1. The EP is a patent system that grants patents for a number of independent countries that are member states of the European Patent Convention (EPC). At present, the European Patent Convention is valid for Austria, Belgium, Bulgaria, Denmark, Germany, Estonia, Finland, France, Greece, Ireland, Italy, Liechtenstein, Luxembourg, Monaco, The Netherlands, Portugal, Rumania, Sweden, Switzerland, Slovak Republic, Slovenia, Spain, Czech Republic, Turkey, Hungary, United Kingdom, and Cyprus. The European Patent Office grants patents which are afterwards administered by the contracting states chosen by the patent owner.
2. *The Regents of the University of California v. Eli Lilly and Co.* 119 F.3d 1559 (Fed. Cir. **1997**).
3. http://www.nywd.uscourts.gov/decision/20030305_00cv6161_larimer.pdf
4. *Burroughs Wellcome Co. V. Barr Labs., Inc.* 40 F.3d 1223, 1228 (Fed. Cir. **1994**).
5. http://wipo.int/pct
6. http://www.uspto.gov/
7. Jaenichen et al. *From Clones to Claims*, pp. 541-544, 3rd edition, Carl Heymanns, Cologne, 2002.
8. R.B. Hildreth *Patent Law: A Practitioner's Guide*, 3rd edition incorporating release no. 4, October **2002**, Practicing Law Institute, New York.

Subject Index

A
adhesin 32
affinity
– capture 47
– chromatography 72
algorithm
– doping 2
– genetic 144
– patenting 192
aptamers 1, 66f, 75
– characterization 77
– DNA 68, 76
– RNA 66, 68
– selection 72
assay
– cellulose filter binding 78
– gel-shift 79
asymmetric synthesis 114

B
binders 51
binding
– cellulose filter 74
biocatalysts 114
– enantioselective 2
– industrial 2

C
cassettes
– random 6
catalytic nucleic acids 87, 102
cell sorting
– fluorescence-activated 31f
– magnetic 33
chirality *see* asymmetric synthesis
codons
– stop 6

D
deoxyribozymes 66
directed evolution 143
display
– cell surface 31
– level 55
– phage 31, 47, 52
– surface 32, 34
diversity 2, 5, 13, 15
– functional 15
– molecular 1
DNA
– catalytic 87
– synthesis 143
doping 143

E
enantioselectivity *see* asymmetric synthesis
error
– frequency 7
– rate 7
evolution
– applied 1
– directed 1, 5, 25, 29, 47
– natural 13

F
FACS 1, 35, 37, 40, 42f
FACS screening 31
filtration
– cellulose 72
folding
– RNA 3
FTIR spectroscopy 125
function 31
– functional 2

Evolutionary Methods in Biotechnology. Edited by S. Brakmann and A. Schwienhorst

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

ISBN: 3-527-30799-0

G
gel-shift assays 72
genotype 31

H
high-throughput screening 114
– assays 133
– high-throughput *see* assays 114
HIV
– reverse transcriptase 5
HTS *see* high-throughput screening

I
immunofluorescence 36
immunoprecipitation 72
in vitro evolution 89
inhibitors
– mechanism-based 47
intimin 32, 34
intramers 68

K
Kazlauskas test 130, 138
kinase
– HSV-1 thymidine 5
knowledge-based potentials 156
Kohonen map 144

L
labeling
– active-site 55
library 47f
– combinatorial 1, 31, 40
– construction 37, 55
– diversity 57f
– first-generation 58
– gene 31
– low-affinity 82
– matrix-binding 82
– molecular 1, 5
– nucleic acid 66
– peptide 31
– phage-display 51
– production 58
– protein 31
– random 87
– – design 91
– – preparation 92
– screening 40, 43
ligands 47
local optimum 144

M
MACS 33, 35, 37, 40, 42f
mass spectrometry 117
microarrays 134
misincorporation 7
mutagenesis *see also* PCR 153
– combinatorial 2
– *in silico* 2
– point 6
– random 1, 5
mutagens 5
mutation
– deletions 8
– frameshift 8
– frequency 8, 10
– insertions 8
– point 15, 25, 29
– rate 5, 29
– transition 8
– transversion 8
– type 5, 8

N
NMR spectroscopy 121
nucleic acid
– catalytic 1, 87
nucleotide
– analogs 5

O
oligonucleotide 56
– degenerate 56
– spiked 6
– synthetic 6
cloning
– degenerate oligonucleotides 57

P
patenting 3, 192
– algorithms 192
– enablement 197
– inventiveness 196
– novelty 193
– priority right 195
– search tools 204
– USA 206
– utility 197
– written description 197

PCR 14, 25, 66
– amplification 10, 28
– background 28
– degenerate primers 55f
– error-prone 5, 8, 48, 55f
– hypermutagenic 5
– mutagenesis 7
– mutagenic 10, 95
– primerless 13
– product 28
– purification 28
phage
– concentration 54
– display 47, 51
– elution 52
– helper 51
– mutant 47
– pIII 51
– pVI 51
– pVII 51
– pVIII 51
– pIX 51
– titers 54, 62
phage-enzyme 48, 51, 63
– activity 54
– capture 60
– eltion 60
– extrusion 53
– purification 53
phenotype 31
phenotype/genotype coupling 31
polymerase
– DNA 25, 27
– Pfu 25
– proofreading 25
– Taq 5, 26f
– Vent 25ff
pool
– DNA 70
– genomic DNA 70
– randomized regions 70
– RNA 70
– starting 69
– synthetic 70
preselection 66, 68, 73
protection
– legal 3
protein engineering 143, 154
protein folding 153
protein structure analysis 153

R
random
– fragmentation 13, 15, 17
– sequence pool *see also* library, random 87,
– – doped libraries 143
randomization 6
randomized region 69
reassembly 16, 19
recombination 10, 13
– DNA 25
– in-vitro 1, 25, 29
– natural 25
repertoire, molecular *see also* library, random 87
reverse transcription 66
ribozyme *see also* nucleic acids, catalytic 66, 87
RNA *see also* nucleic acids, catalytic
– catalytic 87
– ligation 99
RNA
– 2′-modified 75
– pool 66
– purification 66
RNA
– consensus structures 183
– folding 178
– purification 71
– secondary structure prediction 178, 181

S
screening 1, 31, 63
– high-throughput 1
selection 1, 31, 58f, 65, 68, 105
– catalytic elution 62
– in vitro 58, 66
– labeling 59
– main 73
– product labeling 62
– SELEX 186, 192
– strategies 61
– suicide substrate 59
– using a suicide leaving group 61
SELEX 66, 69f, 186
sequence
– complementarity 25
– context 8
shuffling 1, 55
– DNA 13f, 22, 48, 56

– family 15
spectroscopy
– fluorescence correlation 78
– surface plasmon resonance 77
StEP 1, 25, 29
– efficiency 29
strain 8
– mutator 5f, 8
– non-mutator 10
– XL1-Red 6, 8f
substitutions 5
substrates
– selection 48
– suicide 47f
surface 47
– cell 35
– exposure 36

T
T7 RNA polymerase 97
transition-state analogs 47

U
UV/Visible spectroscopy 129

Z
z-scores 154, 160